nature

www.nature.com/nature

Vol 453 | Issue no. 7199 | 26 June 2008

The unlikely matters

The study of cosmic impacts and the effects they have offers two lessons for students of science.

On 15 June, a telescope near Tucson, Arizona, picked up a relatively fast-moving dot in the sky: an asteroid of the sort the telescope is designed to seek. Now known as 2008 LV16, this small rock's orbit takes it from well out beyond Mars to almost as far sunward as Venus, passing by Earth on the way. LV16 poses no threat to anyone, and probably never will. But there is always a chance that, way down the line, something will get in the way of its wanderings. If that obstruction is the Earth, LV16 will cause an explosion on a par with an all-out nuclear exchange between superpowers.

What is remarkable about LV16 is not the risk it poses; over the next century that risk is negligible. It is that the risk has been recognized and quantified — as it has for many, if not yet all, potentially hazardous asteroids. As Alan Harris shows in a Commentary on page 1178, part of this week's special package of material on cosmic impacts, the simple expedient of finding the larger near-Earth asteroids and analysing their orbits has reduced the best estimate of the danger they pose to human life by about 90%.

When cataloguing near-Earth asteroids was first discussed in the 1980s, much of the astronomy community was non-plussed, and some of it was actively hostile. A fair sample of these bodies' orbits, it was argued, is all a scientist needs to see: listing them all would be stamp-collecting, not science. And from a strictly scientific perspective, this had some merit. But as a response to a real, if rather goofy-sounding, threat, it was not enough. The creation of an exhaustive catalogue was the right thing to do.

This illustrates a more general principle: the public-interest response to a problem uncovered by scientific research is not necessarily to do more scientific research. Something more structured and pragmatic may be appropriate, even if it is less likely to bring academic kudos. When the public funds research, it expects, when the need arises, that scientists will use their knowledge and tools for more than just their own scientific interests, even if that imposes some opportunity costs on their discipline.

Given the success that sky surveys have had in reducing risk, the natural impulse is to extend them to ever-larger numbers of ever-smaller rocks. The actuarial case for doing this is not strong; smaller bodies present a much lower cumulative threat than their larger brethren. But although heroic efforts to extend the surveys in this way would be money ill spent, continued progress at modest cost would reflect a certain political shrewdness. Imagine an unspotted asteroid laying waste to a significant chunk of land, as happened in the Tunguska region of Siberia 100 years ago this week; and imagine if that area, unlike Tunguska and a surprising amount of the globe today, were populated. The politician or scientific adviser who had dismissed such a disaster as being too improbable to bother with would be in dire straits. Politicians know in their bones that unlikely events matter.

And that is the second lesson that the study of cosmic impacts teaches. Although impacts follow statistical patterns, individual events and their quirks can have unpredictable effects. As various researchers report in this issue, an impact more than four billion years ago shaped Mars's geological history by dividing its surface into highlands and lowlands, providing a sink in which the planet's water, such as it may have, could pool. An even larger impact gave Earth its Moon. And yet another, at the end of the Cretaceous period, removed Earth's remaining dinosaurs, among a great many other life-forms. Thoughts of impacts have, since the 1980 paper in which Luis Alvarez and his son Walter and their colleagues first detailed that end-Cretaceous event, re-established the possibility of catastrophic change in the scientific imagination, even when the crisis has nothing to do with asteroids or comets.

The Universe is lumpy, irregular, surprising, dramatic. The better this is understood, the better the interplay of chance and necessity can be appreciated. And that same understanding can, more pragmatically, form the basis for vigilance, for planning and for action. ■

For podcast, video material and more, see www.nature.com/news/specials/cosmicimpacts/index.html.

NASA/JPL/UNIV. ARIZONA (SEE PAGE 1173)

Unbalanced portfolio

British research councils should still foster basic science.

Researchers may believe in science for science's sake, but governments often have different ideas. They consider it their duty to seek return from the tax monies they spend — a point of view that is reasonable and responsible for someone in charge of public funds.

The trick, of course, is to avoid taking too narrow a view of what constitutes a return. In their efforts to be business-like, government funding officers will often try to measure success with corporate-style metrics and milestones. This may work well for some areas of government endeavour, but basic research is not one of them. Almost by definition, the frontier of human knowledge is a realm that has no milestones and that encompasses many dead-ends and failures for every advance. Viewed purely by the numbers, researchers' efforts can seem grossly inefficient.

Some recent developments in the United Kingdom point to the dangers that can arise from this cultural divide. A Special Report on page 1150 describes how government officials, understandably eager for a return on their investment in science, are encouraging research councils to build partnerships with industry, and are redirecting funds towards societal problems such as ageing and climate change.

Such initiatives are a necessary part of any nation's science policy. Indeed, many of the research councils' chief executives, who are perhaps eager to win more money for their programmes, have willingly gone along with them. The challenge is to strike an appropriate balance. In practice, continued pressures have led some councils to cut their basic-science portfolios. They have trimmed investigator-led grants, and slashed funding for fundamental fields such as astronomy and high-energy physics in favour of innovation campuses and government initiatives.

"The trick is to avoid taking too narrow a view of what constitutes a return."

Where adequate funding has not been supplied, the emergent effect of the pressures from government is tantamount to an attack by abandoning basic science. If unchecked, this neglect will lead to the loss of scientific subdisciplines and a decline in such intangible benefits as inspiring the young and national pride. And the pressures on research councils may get tougher, as historical declines in science spending within government departments also need to be reversed.

The person responsible for developing advocacy for research council budgets is the director general of science and research, currently absent within government. When Adrian Smith, a statistician currently principal of Queen Mary, University of London, takes up the job in September, he should make it a top priority to ensure that the government fully appreciates the added value of basic science and the costs of its neglect. ■

Comédie-Française

Regional and minority languages should be protected, in France, and elsewhere.

Quelle horreur ! The 40 élite members of the *Académie française* are jumping out of their *fauteuils*, incensed that legislation passed by France's National Assembly would put regional languages such as Breton, Occitan, Corse, Alsatian, Catalan and Basque into the constitution as part of the national heritage. The members are particularly outraged that the regional languages would get a mention in the first article of the constitution — which defines France as an "indivisible, lay, democratic and social republic" — ahead of the second article, which designates French as the official language. The academy, created in 1635 to guard the purity of the French language, voted unanimously this month to condemn the move as "defying logic", and being a threat to the nation.

Actually, "defying logic", is an apt description of the vote itself. Globalization is already threatening to extinguish half the world's 6,000–7,000 languages. That would be a tragic loss to humanity and our understanding of it, if only because knowledge and culture are inescapably interwined with the languages within which they evolved. Languages also enrich each other, and provide a trove of data for research in linguistics and history. The other main French academy, the *Académie des Sciences*, should make itself heard on the matter.

Multilingualism has other practical benefits. French scientists who speak regional languages in addition to the national tongue testify that early bilingualism has helped them go on to master English and other languages. Some even argue that the thought processes involved have helped them to be better and more creative scientists.

The *Académie française* argues that France's regional languages are so obviously part of its heritage that there is no need for constitutional safeguards. That is disingenuous. It is precisely the lack of constitutional recognition that has blocked France from ratifying key international treaties to conserve minority languages: the courts have ruled that ratification is forbidden by existing constitutional principles, such as the indivisibility of the Republic and the unity of the French people.

Indeed, if earlier French governments had had their way, Breton, which is spoken in Brittany, would have been eradicated long ago. Only stubborn Breton persistence has prevented this from happening, notably through the creation of the Diwan Breton-language schools from the 1970s onwards.

Yec'hed mat (to your health) to that — because regional and minority languages, like endangered species, merit protection. Languages that aren't revitalized through constant exercise die out. It's hypocritical that France, which is one of the first to staunchly defend its own elegant national language, should deny that same right to regions that wish to keep their own languages alive and vibrant. The National Assembly's legislation was rejected last week by France's conservative Senate. But it could yet be reintroduced, and should be: for the sake of both science and its own rich heritage, France should remove the constitutional obstacles as quickly as possible, and ratify the European Charter for Regional or Minority Languages. ■

RESEARCH HIGHLIGHTS

Spotted!

Behav. Ecol. Sociobiol. doi:10.1007/s00265-008-0607-3 (2008)

Eyespots on the wings of butterflies and moths (such as the emperor moth, pictured) are thought to scare predators such as woodland birds. Alternatively, the spots may deflect attention away from the central part of the insects' bodies. But when researchers made model moths and distributed them around Madingley Woods near Cambridge, UK, some of the fake moths with eyespots attracted more predators than those without.

Martin Stevens and his colleagues at the University of Cambridge pinned greyscale paper 'moths' with dark, light or no eyespots on their wings to ash and oak trees. The wings were either obvious shades of grey or the same shade as the bark behind them, and were placed over a dead mealworm to provide a reward for predators.

Eyespots proved costly to those targets that were otherwise well-camouflaged, which suggests that eyespots may evolve more easily in already conspicuous species. The markings were previously thought to be merely less advantageous — not costly — in suboptimal circumstances.

K. TAYLOR/NATUREPL.COM

STATISTICS

Who's the driver?

Phys. Rev. Lett. **100,** 234101 (2008)

Faced with a correlation between two variables — chickens and eggs, say — how do you know which is causing the other? Questions such as this, common to fields as disparate as climatology and physiology, are typically unravelled with a statistical technique called Granger causality. But Guido Nolte of the Fraunhofer FIRST institute in Berlin, Germany, and his co-workers say that this method can falsely attribute causality.

They describe a new technique that relies on how phase differences between the driving and dependent variables change with the frequency of their fluctuations. This proves more reliable when the data contain a lot of noise. The authors use their discovery to identify the order in which certain brain regions stimulate others when human subjects shift from a relaxed to an alert state.

EVOLUTION

Model lives

PLoS One **3,** e2282 (2008)

In evolutionary terms, homosexuality might be a detrimental trait that stops people from passing on their genes. But population geneticists have successfully modelled several theoretical explanations for its maintenance. A trio in Italy has now come up with a simple model that involves just two genes.

At least one of these genes, say the University of Padua's Andrea Camperio Ciani and his co-workers, must be on the X chromosome and act to increase fitness when expressed in females. These conditions produce a population within which a small proportion of individuals is gay and in which this proportion remains stable over time.

ALAMY

MATERIALS SCIENCE

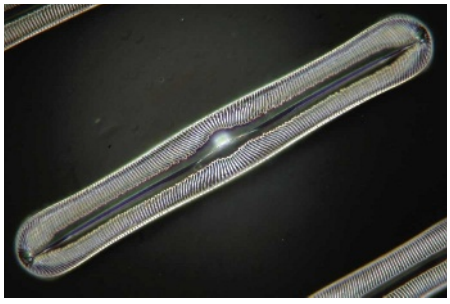
Diatomic power

Adv. Mater. doi:10.1002/adma.200800292 (2008)

Single-celled plankton that have been duped into doping their silica-based shells with germanium can be incorporated into semiconductor chips and made to glow.

Gregory Rorrer of Oregon State University in Corvallis and his colleagues report that the siliceous shells from diatoms (*Pinnularia* sp., pictured below) that were grown for some of their lives in a germanium-rich solution can be incorporated into the devices. On application of an electric field, the shells emit light.

The researchers found resonant frequencies in these emissions that they explain by the geometry of the shells' latticework of pores. They hope that further combinations of semiconductor technology and biologically produced nanostructures may yield novel devices.



MICROBIOLOGY

Infection injection

Science **320,** 1651–1654 (2008).

Two types of disease-causing bacterium use a special injection system to deliver proteins into host cells, researchers have found. The proteins involved contain regions known as 'Anks' (ankyrin repeat homology domains), which often form scaffolds that enable other proteins to interact.

Craig Roy and his colleagues at Yale University School of Medicine in New Haven, Connecticut, report that *Legionella pneumophila*, the bacterium that causes Legionnaires' disease, injects four Ank proteins into mammalian cells via a complex called a 'type IV secretion system'. *Coxiella burnetti*, which causes Q fever, injects eight such proteins.

One of the *L. pneumophila* proteins, AnkX, prevents host vesicles — bags of membrane-bound fluid that contain the bacteria — from moving towards the lysosome, where the bacteria would be destroyed. This may contribute to the bacterium's virulence.

GEOSCIENCE

The geyser forecast

Geology **36,** 451–454 (2008)

The discharge rates of four geysers in North America's Yellowstone National Park — Old Faithful, Daisy, Aurum and Depression — reflect precipitation in the watershed of the Madison River. For decades geologists have tried to link geyser eruptions to external forces including atmospheric pressure and the tides, but until now had little success.

Shaul Hurwitz of the US Geological Survey in Menlo Park, California, and his team propose that years with lots of rain (or snow) result in higher than usual pressures in the 200 °C-plus underground reservoirs that feed geysers, shortening eruption cycles. Conversely, after snowmelt in spring, cold water percolates into geyser conduits and lowers their temperature, lengthening eruption intervals. The work is based on data recorded between 1998 and 2006 by temperature sensors in geyser outflow channels.

GENETICS

The Mod Squad

Nature Genet. doi:10.1038/ng.154 (2008)

The sections of its DNA that a cell expresses — and thus the cell's characteristics — depend in part on chemical modifications to the histone proteins around which DNA is wound. A set of 17 such modifications is associated with a quarter of all promoters — gene-regulatory sites — in human immune cells, find Keji Zhao of the US National Institutes of Health in Bethesda, Maryland, and his colleagues.

They looked at the different combinations of chemical alterations that affect the expression of about 12,500 genes in $CD4^+$ T cells. One type of modification — acetylation — does not directly determine whether a gene is 'read', as had been suspected, but seems to prime the gene for activation.

PHYSIOLOGY

Environmental awareness

J. Gen. Physiol. **131,** 605–616 (2008)

Researchers at Stanford University in California report that proteins can alter the lipid environment around a transmembrane ion channel in a way that influences whether the channel is open or closed.

Miriam Goodman and her colleagues studied an ion channel involved in the nematode worm *Caenorhabditis elegans*'s sense of touch. The channel's core comprises the proteins MEC-4 and MEC-10. Around this core sit two other proteins, called MEC-2 and MEC-6, which are responsible for the observed effect.

MEC-2 is found on the inner side of cell membranes and seems to bind cholesterol. MEC-6 resides on the extracellular side of ion channels and contains a helical structure that may associate with various lipids. It is not clear whether MEC-2 directs the ion channel to regions in the membrane flush with cholesterol or whether it attracts cholesterol to the area around the pore.

CELL BIOLOGY

Motor control

Science **320,** 1636–1638 (2008)

A protein that allows the soil bacterium *Bacillus subtilis* to quickly halt its propeller-like propulsion and thus stick to a surface has been identified by Daniel Kearns of Indiana University in Bloomington and his colleagues. EpsE, the protein, seems to act like a clutch rather than a brake; it leaves the rotors that drive the bacterium's flagella unpowered but spinning freely rather than slowing them down.

The authors labelled *B. subtilis*'s EpsE with a fluorescent protein and revealed that EpsE is associated with flagellar motors. They then attached bacteria to a surface by a flagellum. The cells rotated passively even when they produced EpsE, which would have prevented them from swimming.

MICROFLUIDICS

Groove train

Nature Mater. doi:10.1038/nmat2208 (2008)

A network of grooves only a few millimetres long has been used to guide tiny structures around fluid-filled channels. These structures organize themselves into complex arrangements, as illustrated by the Greek temple pictured below.

SU EUN CHUNG *ET AL.*

More than 50 microstructures were slotted into the grooves. Water flow pushed them to the end of the lines, where they latched into place. Then exposure to ultraviolet light fused them together. The microstructures can carry living cells, among other things, suggesting an application in tissue assembly.

Sunghoon Kwon's team at Seoul National University in South Korea say that this improves on other methods of sub-200-micrometre robotic assembly, which are dearer and slower.

JOURNAL CLUB

Seth Lloyd
Massachusetts Institute of Technology, Cambridge.

A quantum mechanic considers how we might 'talk' to aliens.

So it finally happens. After hundreds of years of humans attempting to communicate with extraterrestrial beings, our descendants receive a message back. But it looks like utter gibberish. What to do? Earthlings might, for example, find some middle ground by sending the aliens a stream of circularly polarized photons to explain what we mean by left handedness. Or maybe the aliens would be able to decipher simple mathematical formulae, encoded in a binary alphabet, through which we could gradually build up a mutual understanding of mathematics, logic, and so forth?

That might work, but what if the replies are still nonsensical? Brendan Juba and Madhu Sudan recently supplied a mathematically precise answer to this question (B. Juba and M. Sudan *Symp. Theor. Comput.* 123–132; May 2008). Using the theory of interactive proofs, which shows how parties who possess different pieces of a theorem's proof can cooperate to construct a full proof, they show that as long as aliens are not completely indifferent to communications from Earth, we will quite quickly be able to ascertain whether or not they have knowledge that is useful to us.

The technique that Earthlings should use goes like this: Bob, the human, systematically encodes questions about a class of problems in a form that any computer can interpret. He then repeatedly sends the encoded questions to Alice, the alien, and carefully parses the apparent gobbledygook that she sends back. Juba and Sudan prove that if Alice knows the answers to Bob's questions (that is, were the questions asked in her own language), and actually answers some non-negligible fraction of those questions (again, in her own language), Bob can determine what she means.

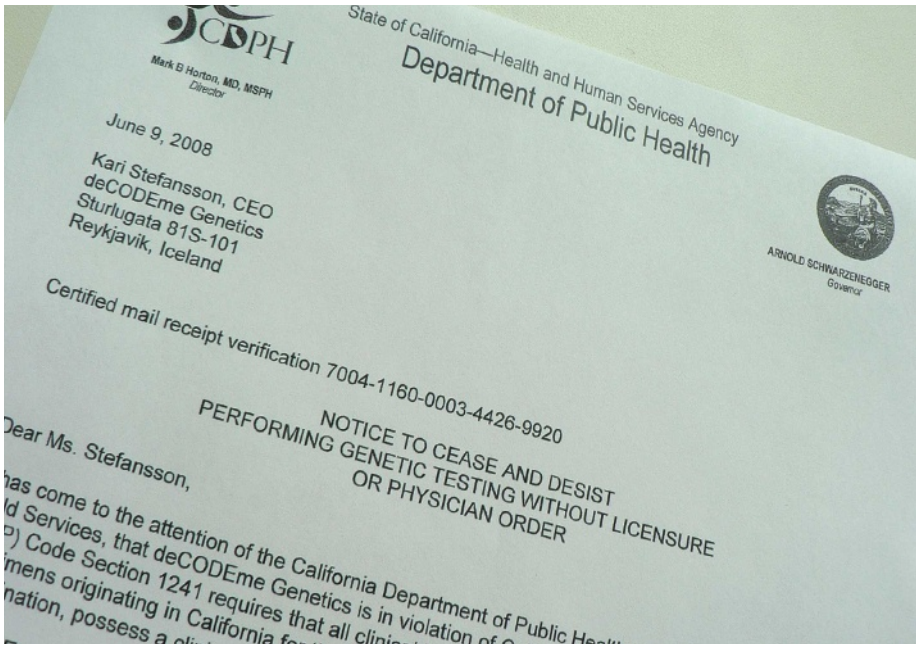
So communicating with aliens is possible in principle, no matter how unpromising the task may seem. I find that reassuring.

Discuss this paper at **http://blogs.nature.com/nature/journalclub**

NEWS

Gene-testing firms face legal battle

The state of California is clamping down on companies that offer direct-to-consumer genetic testing in a move that threatens the burgeoning industry. **Meredith Wadman** looks at a grey area in US regulation.



Last Wednesday, as California governor Arnold Schwarzenegger prepared to tell a biotechnology industry convention in San Diego that his state "is one of the best places to set up shop", Kári Stefansson was opening a letter that had just landed on his desk at deCODE genetics in Reykjavik, Iceland.

The letter read: "It has come to the attention of the California Department of Public Health...that deCODEme Genetics is in violation of California law" for failing to have a clinical laboratory licence in the state and offering genetic tests to consumers resident in the state without a physician's order. It gave deCODE until 23 June to submit a plan showing how it would correct the situation, or face "civil and/or criminal sanctions".

Stefansson's high-profile company is one of 13 genetic-testing firms that have been targeted during the past two weeks by the California agency with a letter to "cease and desist" selling tests to California's residents. The directive poses a serious challenge to plans for a new era of Internet-based, direct-to-consumer genetic testing. The companies include Californian businesses 23andMe and Navigenics, which have begun marketing test packages based on genome-wide arrays within the past six months, and DNATraits.com, based in Houston, Texas, which counsels prospective parents on the genetic risks faced by their future offspring.

DNA Direct, a San Francisco-based firm founded in 2003 that offers tests for familiar mutations in well-characterized genes such as *BRCA1* and *BRCA2*, which convey an increased risk of breast cancer, did not receive a letter.

"If these companies were constrained from reaching out to consumers directly, it would certainly cause some near-term difficulty for their business," says life-sciences analyst John Sullivan at Leerink Swann, an investment bank based in Boston.

During a public meeting on 13 June, Karen Nickel, California's chief of laboratory field services, who wrote the letter, said that consumer complaints had triggered an investigation into 25 companies, of which 13 were ultimately sent letters. Nickel told the meeting: "We [are] no longer tolerating direct-to-consumer genetic testing in California." Under California law, the companies could be fined up to US$3,000 per day for each violation if there is no "immediate jeopardy" to state residents — and from $3,050 to $10,000 per day if there is.

23andMe declined an interview request. It released a statement emphasizing that it is an "informational service", and said it is "eager to work with" regulators in California and elsewhere to develop appropriate regulations to govern the nascent industry.

"It doesn't increase your risk. It measures your risk."

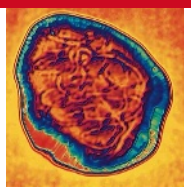
But two of the companies dispute the charges in the letter. Mari Baker, chief executive of Navigenics, based in Redwood Shores, says that its tests are read by a licensed, certified laboratory and that a company physician is involved both in the approval of a genetic-test order and when the results are released to a customer. "It's important to do this the right way," says Baker. "And that we are doing. So this has come as quite a surprise. The only conclusion we can come to is we have not properly informed the state as to all the steps we have in place. We have reached out to them to try to schedule such a meeting."

Stefansson says that deCODE is not marketing to California residents; its website lists California among several states for which the company's deCODEme Genetic Scan "may omit certain information" because of state law. He says that his company is in the process of trying to obtain a California licence, but that the state has been unresponsive. He is a physician and oversees the ordering process for deCODE, he says, adding that a large proportion of the test requests the company receives are from people's personal physicians.

Even so, he calls the legality of California's requirement that a physician order a genetic test "questionable". He says: "I don't think that they can basically tell the people of California that 'you cannot order a test like this for yourself without going to a physician'. I don't think they can raise that kind of barrier." He argues that a genetic test is not a medical intervention like, say, a prescription for a cholesterol-lowering drug. "It doesn't increase your risk. It doesn't decrease your risk. It measures your risk. It's a description of who you are."

The state Department of Public Health — which is no longer making Nickel available for interviews — declined to say whether a company-employed physician who oversees orders initiated by customers through the Internet would meet its requirement that a physician order the genetic test.

That has been a grey area among regulators and policy-makers, says Kathy Hudson, director of the Genetics and Public Policy Center at Johns Hopkins University in Washington DC. "Is a doctor who is your personal physician, who has a fiduciary responsibility to you,



MEASLES MYSTERY
The disease doesn't work as was thought.
www.nature.com/news/

A. PASIEKA/SPL

the same as a company doctor? Is it really the same thing to call up a company who is trying to sell you a product and have their doctor, who receives a pay-cheque from them, advise you? To me, it is not the same thing."

The Californian law does not prevent state-based companies from marketing the tests to consumers residing outside the state. But a muddle of different state laws across the United States provides a confusing array of regulations on the issue.

New York's Department of Public Health has sent similar letters to 26 companies since last November. In response, 23andme and Navigenics have submitted business plans that the department is reviewing, says its spokesman Jeffrey Hammond. "The good news for both 23andme and Navigenics is the labs they are proposing they use for their lab work are known to the department and have a history of compliance with us." Hammond adds: "Our goal is not to levy fines. It is to bring companies who want to do business in New York into compliance with state law."

Hudson says that the California and New York letters point out the conspicuous lack of unifying federal regulation of genetic testing. The law "is very inconsistent from state to state at a time when the risks to consumers do not vary state to state — and when we have businesses that are certainly operating state to state".

Ryan Phelan, chief executive of DNA Direct, says that the one-size-fits-all approach of the letters is a cause for concern, because tests being offered range from those predicting serious diseases to "fun to know" information with no bearing on a person's health. "There is going to be increasingly a need for nuanced regulation. All genetic tests should not be considered the same." ■

Biogen fights takeover bid

The latest chapter in the battle over Biogen Idec, a prominent US biotechnology firm, ended last week when shareholders rejected billionaire investor Carl Icahn's bid to oust three members of the company's board of directors. The decision was a clear defeat for Icahn's plan to gain control of the company, which is based in Cambridge, Massachusetts, and a victory in Biogen's struggle to remain independent.

The fight between Icahn (pictured) and Biogen began last August, when Icahn announced that he had acquired 1% of the company's stock, making him one of its biggest investors. At first, Biogen officials weren't sure what to make of the news, says Naomi Aoki, Biogen's director of public affairs. Icahn is a classic corporate raider, known for gaining control of companies and then forcing them to take actions that increase the value of their stock. Over the past 30 years, Icahn has compelled many of his investments to sell their assets altogether.

C. EAST/REUTERS

"We knew his reputation," says Aoki. "But I don't think it was clear to us right from the beginning how everything would play out."

Icahn's goals fit cleanly with the drug industry's rising demand for biotechnology companies. In April 2007, pharmaceutical giant AstraZeneca surprised analysts by paying US$15 billion for MedImmune, a Maryland-based biotechnology company. Icahn, a MedImmune shareholder, later acknowledged that he had pressured MedImmune's board to sell, and had threatened a hostile takeover if the company did not comply. Biogen, a 30-year-old company with several candidate drugs in the pipeline, is worth $23 billion.

In October, Biogen announced that it would accept offers from firms interested in buying the company. It called off the hunt for a purchaser in December, saying that no one had come forward. Icahn accused Biogen of interfering with the search, sued the company for access to records pertaining to the failed sales process, and announced that he was nominating three candidates to the board of directors. It would have been the first step in a two-part plan to gain control of the board that would have been completed next year when four more seats on the 12-member board became available for re-election.

In a memo to shareholders issued two weeks before the vote, Icahn accused Biogen of lying to them about its attempts to sell the company. Biogen issued a letter of its own, saying in bold-faced capital letters: "Do not be misled by Carl Icahn."

On 19 June, a preliminary tally of shareholder votes showed that Biogen had won this round. Icahn has not said what his next move will be, and was not available for comment. But he now controls 4% of the stock and has said that he has some of the top ten investors on his side, suggesting that Biogen's battles may not be over. ■

Heidi Ledford

Scientists get online news aggregator

A Canadian graduate student dissatisfied with science coverage on online sites such as Google News and Yahoo News has created a news aggregator especially for scientists.

Michael Imbeault, an HIV researcher at the Université Laval in Quebec, launched his fully automated site called e! Science News (http://esciencenews.com) last month. It has already attracted 300,000 different users, and averages 5,000 visits a day, he says.

News aggregators display headlines and snippets from other media sources, but don't produce their own content. Of the top five online US news sites, three are aggregators — Google News, AOL News and Yahoo News — and only two — CNN.com and MSNBC.com — generate original content. Yahoo and AOL use human editors and source almost all science stories from wire agencies, such as Reuters. Google News uses computer algorithms to aggregate headlines from thousands of news sources, ranking them by how often and on which sites stories appear. Science and technology coverage on Google News, for example, is notoriously devoid of basic science.

Imbeault's site indexes science news sites, clusters similar articles together on the basis of the frequency of word co-occurrence, and then uses Bayesian statistics to automatically assign articles to topics such as astronomy, health and climate. It then ranks them using factors such as timeliness, and the number of sites reporting the same news, which indicates the story's importance. At present, it is limited to around 40 news sources — including *Nature News*, *The New York Times* science section and institutional news sites such as NASA, which offer free content for at least a period — but this will be increased, he says.

Imbeault built the site on top of the Drupal open-source content management software. He says that his aggregator will also be improved by moving to semantics-based techniques that better capture the meaning of a text. ■

Declan Butler

SPECIAL REPORT

Payback time

The UK government has invested heavily in science. Now it's looking for a return, and some worry that the research councils are being pressured to deliver, possibly at the expense of 'blue skies' research. **Geoff Brumfiel** looks at the changing landscape of science funding in Britain.

British scientists have had it pretty good this past decade. Since 1998, government funding for the nation's seven research councils has nearly doubled in real terms to around £3 billion (US$5.9 billion) this year. The boost has helped make the United Kingdom an attractive country for science.

But there is growing concern in the research community that the increases are coming with a cost. Some believe that the Treasury is using its influence to erode the independence of the research councils, which fund the majority of basic science in the United Kingdom. Critics point to the latest round of council documents, which are littered with Treasury catchphrases such as "economic competitiveness" and "social impacts".

Some councils have begun favouring government initiatives over investigator-motivated projects, and at one in particular, the Science and Technology Facilities Council (STFC), deep cuts have been made to fundamental fields, in part to protect more commercially appealing programmes.

The signs all point to one thing, says Philip Moriarty, a nanotechnology researcher at the University of Nottingham: the government is using its money to push the country's scientific enterprise towards commercial profitability. "Blue-skies investigator-driven research is getting squeezed out," he says.

It's a bleak view that not everyone shares. "In most respects, I think we're in a fairly healthy situation," says astronomer Martin Rees, president of the Royal Society, the UK's national academy of science. Some strategic changes have been made, he adds, but given the government's heavy investment in research "one should not be surprised". Overall, the councils have continued to maintain independence in their daily operations.

Regardless of their opinion on the issue, UK researchers agree that change is afoot. "I think that the Treasury is increasingly questioning the return on its investment in R&D," says neurobiologist Colin Blakemore, who led the Medical Research Council (MRC) from 2003 to 2007. "The issue is what role the research councils should play in research and innovation."

Ancient principles

Britain's research councils pride themselves on their autonomy. Under a 1918 recommendation known as the 'Haldane principle', governments have abstained from interfering in day-to-day operations. This informal rule has allowed funding to be distributed by a system of independent peer review that is widely credited with making Britain a European leader in research.

In recent years, a steady flow of funding has helped to further strengthen the UK's position. The increases began under Prime Minister Tony Blair, but were driven in large part by the current prime minister, Gordon Brown, when he headed the Treasury as Chancellor of the Exchequer. Brown made investment in science a top priority of the Labour government. A 2004 government policy paper co-authored by Brown outlined his long-term ambitions for scientific investment: research and innovation is "key to improving the country's future wealth-creation prospects," the report said. It called for increased funding for research, but argued that more attention needed to be paid to ensuring that basic science translates into financial benefit for the nation.

The paper's assertions have been echoed by a series of policy briefings and independent studies sponsored by the government. A 2006 report for the government by David Cooksey, a former director of the Bank of England and former governor of the Wellcome Trust, recommended that the MRC increase its investment in translational medicine, the medical application of basic scientific work. In 2007, a second, broader, review by David Sainsbury, a former government science minister, recommended building stronger ties between research and industry.

In October of last year, similar recommendations appeared in the government's comprehensive spending review, which set spending levels from 2008 to 2011. Most strikingly, the Treasury awarded the MRC a £160 million boost to its existing £540 million budget. Most of the increase, £130 million, will go towards translational medicine.

Many viewed the new money for translational work in a positive light. "I think it's completely fine," says Harpal Kumar, the chief executive of Cancer Research UK, Britain's largest

T. PILSTON/THE TIMES/NPA ROTA/PA

Prime Minister Gordon Brown advocates research that can be translated into societal benefits.

Y. ARTHUS-BERTRAND/CORBIS

The UK Treasury's influence may be reducing the amount of fundamental research done in the country.

cancer charity. "I don't think it would get done if the government didn't push in that direction." But this translational push is uncontroversial because it comes from new money, not basic research funds.

At other councils, where increases keep pace with inflation, the leadership is choosing programmes that align with Treasury priorities. The Natural Environment Research Council (NERC) is emphasizing a new, government-led climate-change initiative, while requiring researchers seeking undirected funding to explain the wider, societal benefit of their studies. The Engineering and Physical Sciences Research Council, which oversees the majority of materials science and physics grants, has cut back its funding for investigator-motivated grants by 15% to pay for several government initiatives. In addition, all councils are pledging a portion of their funding to a newly elevated Technology Strategy Board, which will seek to build industry–academia partnerships.

Again, some dispute whether the changes are significant. For example, NERC already had a heavy interest in climate change, says Alan Thorpe, the council's chief executive. The government's climate-change initiative "doesn't stop us from addressing frontier challenges," he says.

But critics worry that change is happening without public debate. The redirection of the councils comes mostly from chief executives, without much public engagement, says Phil Willis, head of the House of Commons select committee on innovation, universities and skills, which scrutinizes the councils. "I think it is happening by stealth", he warns. "These are worrying trends that need to be addressed."

The one council where subtlety is lacking — and where the Treasury's influence may be most damaging — is the STFC, which oversees large-facility physics and astronomy. Faced with a shortfall in its roughly £573 million budget, the council announced last year that it would cut by a quarter grants for high-energy physics and astronomy. In addition, it has withdrawn from the International Linear Collider, the world's only next-generation particle accelerator project. The cuts have helped to pay for new X-ray and neutron sources which can be used for both academic and commercial purposes. Those facilities are integral to creating the council's new Harwell Science and Innovation Campus in Oxfordshire, according to planning documents.

"Blue-skies investigator-driven research is getting squeezed out."

Shortly after the announcement, Keith Mason, the chief executive of the council, told researchers working in the more fundamental fields funded by STFC that they needed to work harder to prove their economic worth.

This message left many physicists angry and disappointed. The societal payback from fundamental physics comes mainly from well-educated graduates for the workforce, says Brian Cox, a high-energy physicist at the University of Manchester. "The government has damaged what to me is our core business as a research council," he says. Moriarty goes further: "The cynical point of view is that it's part-and-parcel of New Labour's ethos to transfer as much money to industry as possible," he says. The goal, he says, is to make university researchers "contract workers for industry".

Free to choose

Those accusations are categorically denied by John Denham, the secretary of state for innovation, universities and skills, who oversees the councils. "We've not sent instructions to the research councils to say that you choose your funding according to whom can show economic return," he says. In fact, the government has announced a full review of STFC management in response to the outcry. Nevertheless, the councils have been "encouraged" to show where they have created an economic benefit, he says. Such returns will "strengthen my argument with the Treasury to get more money".

It is also true, Denham adds, that the government has asked the councils to direct funding towards certain specific areas such as ageing and climate change. "There are some strategic, massive problems that we as a society face. We want to be sure that part of the research efforts maximizes our chances of dealing with those problems."

Blakemore agrees that the government and the Treasury are not forcing changes through the councils. He believes that much of what's happening is driven by council leadership, who make changes in hopes of winning more funding. "I think, frankly, the increase at the MRC was seen as an experiment really," he says. And as long as funding increases accompany the changes at other councils, he adds. "I think it's likely to be win-win."

But others are sceptical. The councils are most valuable as backers of fundamental science and should be kept free from commercial interests, says Moriarty. "The government should be funding research that the markets won't."

Ultimately, in the current difficult economic environment, Willis believes that continued pressure will cause funding for investigator-driven fundamental projects to shrink. "The worst-case scenario is that you go to what I would call a command economy science," he says. "I think that would spell the death-knell of British research." ■

See Editorial, page 1144.

PUNCHSTOCK

CONSERVED LANGUAGE
Conservation aims to get 'truly scientific' terms.
www.nature.com/news/

When water gushed on Mars

Were the northern plains of Mars submerged in a vast flood as recently as 20,000 years ago? Geologists claim to have found evidence of a recent volcanic eruption under the ice cap that could have created a wall of water 200 metres high and 35 kilometres wide.

Signs of volcanic activity and flowing meltwater have been found before, but the new study links the two together with strong geological evidence, bolstering theories that water was the chief sculptor of the huge chasms in the northern martian ice cap. The flood, the researchers say, could have occurred within the past 10 million years and maybe as recently as 20,000 years ago — more evidence that Mars has not been a geological corpse since its wet and warm period billions of years ago.

Conditions in the time leading up to the flood — sulphur-rich volcanic gases, warmed water, an ice cap shielding against ultraviolet light — would have been similar to those at deep-sea vents on Earth that are crucibles for primitive life. "You have a set of conditions that comes close to what people have speculated would be favourable for life," says geomorphologist Niels Hovius at the University of Cambridge, UK, who led the study. "This would be a place where I would start looking."

G. NEUKUM/ESA/DLR/FU BERLIN

Basalt cones at the edge of the martian northern ice cap show evidence of flood scouring (inset).

The geology in this region of Mars is strikingly similar to features in British Columbia and in Iceland, where volcanoes and glaciers interact, says Hovius, who usually studies those environments. He began his martian venture with an image from the European Space Agency's Mars Express orbiter that shows cones of basalt standing in a plain at the edge of the ice cap. Hovius noticed that the cones had material piled up on one side and scoured away on the other — like bridge piers in a river.

"The geomorphic evidence is screaming out that Mars has been active in relatively recent times."

That terrain leads to a chasm in the ice cap and then to a ring 6 kilometres across. Flat-topped ridges extend outwards from the ring. Hovius's interpretation is that the ring is volcanic in origin (it is too deep in regard to its diameter to be an impact crater), and that the ridges are tuya, rock formed when lava erupts into ice. The amount of ice that could be melted by the mapped tuya matches the volume of 'missing' ice in an area of subsidence — in which the chasm lies — of several thousand square kilometres, Hovius says.

The study, which is in press in the journal *Icarus*, adds to the evidence that the many chasms in the ice were started by water and perhaps later expanded by wind, says geologist Kathryn Fishbaugh at the Smithsonian Institution in Washington DC. "It's a strong paper," Fishbaugh says. "I think it's the best you can do for how melting might be triggered."

Fishbaugh says the only controversy about the study is the purported timing of the flood. The age of 10 million years is based on a lack of impact craters in the presumed flood area, but dates based on cratering records are notoriously uncertain. The 20,000-year estimate is even more speculative, being based on a blanket of dust observed on the ice cap, which Hovius attributes to detritus from the eruption. Fishbaugh, on the other hand, doesn't think that the chasm could have been eroded that quickly.

The flooding scenario could be confirmed with better images. Fishbaugh says the region is in the queue of high-resolution pictures to be taken by NASA's Mars Reconnaissance Orbiter by August. Meanwhile, the Phoenix lander, in the first month of its mission, sits in the northern plains on the other side of the ice cap (see *Nature* **453,** 576; 2008). It is scraping through a thin layer of soil to the ice, as it attempts to unravel the area's history. If Phoenix finds evidence for recent melt water, the outburst floods could be a new mechanism providing it, says Victor Baker, a hydrologist at the University of Arizona in Tucson, who has proposed similar Mars flooding scenarios. Other explanations for any changes that Phoenix might find in the ice invoke changes in climate, caused by changes to Mars's tilt and orbit.

Even if Hovius's idea proves correct, there is still the question of what caused the volcanism in a seemingly inactive world. Just last month, researchers using radar on the Mars Reconnaissance Orbiter to peer through the ice cap found that the bedrock beneath was largely undeformed, unlike the warmer outer shell of Earth, which bends slightly under the overlying weight (R. Phillips *et al. Science* **320,** 1182–1185; 2008). If the outer shell of Mars is so stiff and cold — even colder than the Moon, according to this study — how can hot material rise up from its interior? "It's typical Mars," says Baker. "We have a whole lot of things that seem to say that Mars is or should be dead. Yet the geomorphic evidence is screaming out that Mars has been active in relatively recent times." ■

Eric Hand

Population genomics for fruitflies

Starve a fruitfly for a couple of hours, and it gets a little cranky. Pop it into a fruit-fly-sized ring, add seven other starving flies and just one piece of food, and you'll have a riot. "They stand up on their little hind legs and tussle," says geneticist Trudy Mackay at North Carolina State University in Raleigh. "And then there's the wing slap," adds her collaborator, Robert Anholt.

Researchers in Mackay's lab quantify aggressive behaviour by counting how often each fruitfly wrestles, slaps, or chases its competitors. They have uncovered a wide range of responses, even among members of the same species. One fly had 100 aggressive interactions in two minutes; others had as few as three. "Those are the pacifists," says Mackay. "They sit there and share the food." Anholt sums up the data succinctly: "There's enormous variation — from flies that are real wimps to flies that really beat the shit out of each other."

It's that sort of variation that has inspired Mackay and her colleagues to propose an ambitious new drosophila genomics project: sequencing the genomes of 192 flies. The aim is to use this information to understand the genetic changes underlying the variation in behaviour and appearance in natural populations. By breaking free from lab strains that have spent decades living in vials on lumps of smelly, sterilized lab food, researchers hope to get a better glimpse of the evolutionary forces at work in fly populations. "The history of a population can leave a kind of footprint on the genome," says fruit-fly researcher Esteban Hasson of the University of Buenos Aires. "Perhaps we will be able to map the mechanisms that shaped the variation found in these flies."

OXFORD SCIENTIFIC/PHOTOLIBRARY.COM

Researchers want to sequence the genomes of a swarm of the geneticist's favourite fly.

The flies are all members of the same species — the classic genetic model organism *Drosophila melanogaster* — but, unlike some lab populations, they have been kept in the lab for only five years — since their ancestors were trapped at a farmers' market in Raleigh. They have been inbred for 20 generations in the lab to produce a collection of 'pure-bred' flies that are amenable to genetic analysis. Flies obtained from the wild have been used to study natural variation before, but the work can be arduous. Having full genome sequences promises to dramatically speed up the process, says Mackay.

This project, which was recently approved by the US National Human Genome Research Institute in Bethesda, Maryland, is expected to cost around US$4 million and to take about two years. Although similar projects are under way in other model organisms (see 'What's the magic number?'), most of those are smaller and will sequence only targeted regions of the genome. Even when a full genome is the goal, it is typically only a 'draft' sequence that contains a relatively high frequency of gaps and errors.

The drosophila project, in contrast, aims to produce high-quality sequence using two high-throughput sequencing methods with complementary strengths. The machines produced by Illumina of San Diego, California, are particularly good at detecting single base differences in the DNA sequence, whereas the sequencers from 454 Life Sciences of Branford, Connecticut, are better at finding regions in which large stretches of DNA have been inserted or deleted. "I'm after as much sequence as I can get, in a very greedy way," says Stephen Richards of Baylor College of Medicine in Houston, Texas, who is an investigator on the drosophila project.

This approach will provide a useful testing ground, says population geneticist Philip Awadalla of the University of Montreal in Canada. Awadalla usually studies genetics in humans and in the malaria parasite, but says the analytical tools developed in the drosophila project could be useful for his work. Awadalla also plans to analyse the drosophila sequences once they become available. The small genome and the ease with which different strains of *D. melanogaster* can be characterized provide clear advantages over humans or other insects, he says.

What's the magic number?

As the price of large sequencing projects drops, researchers are launching ever more ambitious plans to 'resequence' full genomes. These projects aim to sequence additional individuals from a species that has already had its genome sequenced. For example, it's more than a decade since the genome of the model yeast *Saccharomyces cerevisiae* was published, and since then 36 strains isolated from the wild have been fully sequenced.

But the multicellular model organisms, with their bulky genomes, have not kept pace. Fifteen strains of mice have been resequenced and three wild isolates of the nematode *Caenorhabditis elegans*, with plans to sequence a dozen more. And the full genome sequences of just two people have been published, geneticists Craig Venter and James Watson.

The numbers are set to go much higher. The most famous resequencing endeavour is the 1000 Genomes Project, which aims to sequence the genomes of 1,000 people. Less well known is another project, still in the planning stage, that hopes to literally one-up the 1000 Genomes Project: the proposed 1001 Genomes Project, headed by Detlef Weigel at the University of Tübingen, Germany, to sequence 1,001 different strains of the flowering plant *Arabidopsis thaliana*. In the meantime, Weigel has received funding to sequence 80 strains from around the world. **H.L.**

Drosophila researchers are renowned for devising creative ways of studying their beloved flies. Researchers are already lining up to study Mackay's wild isolates, with plans to study learning and memory, wing morphology, body size, social behaviour, circadian rhythm and responses to different odours and drugs. Hasson will try to determine why some flies like to lay their eggs in grapes whereas others prefer oranges, and Anholt will test responses to alcohol using his 'inebriometer', a device that measures how quickly flies become woozy from ethanol fumes.

Richards says that similar projects in other insects, such as mosquitoes or honeybees, could be on the horizon. "The cost of sequencing is coming down so quickly; in the future it'll just be a normal grant proposal to do 500 insects," he says. ■

Heidi Ledford

Online anthropology draws protest from aboriginal group

As Europe's museums begin archiving their collections in digital format, skeletons are emerging — and not just of the physical variety. One South African tribe already says it will oppose the inclusion of images of its people's remains in any multimedia format.

The University of Vienna has started to digitize the collection made in the early twentieth century by Rudolf Pöch, considered one of anthropology's founding fathers. The project, headed by Maria Teschler-Nicola, will improve the collection's accessibility for researchers and store the delicate material in a sustainable way, using electronic records of physiological measurements as well as two- and three-dimensional scans.

But the full collection, which includes human remains and thousands of ethnographic artefacts, was gathered using unethical methods, such as grave-robbing.

Around the turn of the twentieth century, anthropological adventurers in search of exotic artefacts collected skeletal remains from ethnic groups in Africa, Asia and Australia, and sold them to museums in the West. There, they were often displayed in exhibitions purporting to show the evolution of humans from these supposedly 'primitive' origins.

Museums in Europe and the United States have now stopped displaying the remains of modern humans that were not acquired by donation. But it was not until 1995 that the Natural History Museum Vienna removed an exhibit depicting a Negro man as being below Caucasians on the evolutionary scale of development.

"There are maybe 300 sensitive cases in our collection," says Teschler-Nicola. "We don't want to repeat the same mistakes, but we don't have any guidelines." Such bones can be important research material for archaeoanthropologists, which complicates the museum's decision.

The Natural History Museum in London is also planning to digitally record its entire collection, and has yet to decide what to do about its own contentious human remains. The issue was raised at a meeting organized at the museum in March to survey the opinion of leading international scientists. An internal report from the meeting is thought to endorse continuing scientific study, including digitization, on human remains that may be subject to repatriation. "The decision on how to move forward is yet to be taken," says John Jackson, science-policy coordinator at the museum. "There are constraints on whether those remains should be in the collection — whether it is ethically right has to be considered very carefully."

"When repatriation requests are made there is an expectation that all studies have been done. This is not the case," says Robert Hedges, an archaeologist at the University of Oxford, UK, who attended the meeting. "There is every reason for studying remains that are vulnerable to repatriation. You have to be aware that one is liable to lose information if remains are repatriated."

Roger Chennells, legal adviser to the San Institute, a South African non-governmental organization that campaigns for the repatriation of the aboriginal San people's remains, some of which are in the Pöch collection, told *Nature*. "We have not been consulted, and we do not support any photographic archiving of our people's remains — we are opposed to it," he says.

The University of Vienna and the National History Museum in London both hope to draw up guidelines in the next few weeks. ■

Tony Scully

ON THE RECORD

"We have ICE!!!!! Yes, ICE, *WATER ICE* on Mars! w00t!!! Best day ever!!."

Comment from 19 June on the Mars Phoenix 'twitter' feed, where team members leave updates in the persona of their plucky lander.

SCORECARD

▲ **Runner reaction time** Canadian scientists have found that being close to the starting gun startles runners into a speedier start. The team suggests that Canada's Olympic runners wear cranked-up hearing aids in Beijing to get the best reaction time off the blocks.

▼ **Mail delivery time** A 'slow art' project at Bournemouth University in the United Kingdom uses three snails crawling around a tank to pick up e-mail signals and pass them on. The 'real snail mail' can take months to be delivered.

EUREKA/ALAMY; K. TAYLOR/NATUREPL.COM

NUMBER CRUNCH

£9,999.99 The 'N-prize' cash award for launching "an impossibly small satellite on a ludicrously small budget".

£999.99 The maximum allowed cost of the launch.

9 orbits How far it has to fly.

9.99–19.99 grams The required weight range of the satellite.

0.1–1 kilogram The weight range of the smallest common category of satellites, known as 'picosatellites'. One popular one, CubeSat, costs about £20,000 (US$40,000) per launch.

The prize rules state that "imaginative use of string and chewing gum is encouraged".

Sources: Phoenix Twitter, Edmonton Journal, BBC, www.n-prize.com

SIDELINES

US Congress signals new funds for key science areas

Last-minute budgetary wrangling in the US Congress may have netted science agencies several hundred million dollars more than they had expected for the fiscal year 2008.

As *Nature* went to press, the House of Representatives had passed a supplementary budget bill for this financial year. The US$163-billion bill mainly provides extra money for the war in Iraq, but it also contains some domestic spending, including key research priorities.

It includes an additional $150 million for the National Institutes of Health (NIH) and $62.5 million each for NASA, the National Science Foundation and the Department of Energy's Office of Science. The absolute amounts are relatively small (the NIH, for instance, is a $29-billion agency), but are a response to criticisms that funding of the NIH has remained flat for the past five years and that although physical-sciences research was designated a priority, it has not been funded as such during the past two years.

Boost biosafety funding to cut risks, say UK officials

British labs handling the most dangerous diseases need additional funding to ensure that devastating outbreaks do not ravage the nation, an influential group of politicians has said. The review was triggered by last year's outbreak of foot-and-mouth disease, which was traced to a damaged pipe at the Pirbright laboratory in Surrey, and cost the government £47 million (US$92 million).

A report by the House of Commons' science select committee warns of "shortcomings" in Britain's funding of high-containment facilities, especially in terms of maintenance costs. "This must be rectified to ensure the incident at Pirbright is not repeated," the report says.

The report also warns of a "striking" lack of coordination between the bodies that pay for and run high-containment labs.

D. BEREHULAK/GETTY

Foot-and-mouth disease devastated British farms.

Radar and wind farms should coexist, say advisers

The US government must do more to make radar systems and wind farms compatible, according to a group of independent scientific consultants.

The spinning blades of wind turbines are known to interfere with defence and weather radar systems. As wind farms have proliferated, both the US and UK governments have sought to limit their growth (see *Nature* **428,** 111; 2004).

But in the United States at least, the government would be better off working to develop regulations and new technologies that can ameliorate the problem, according to the JASONs, a group of academic researchers that advises US defence agencies on a range of issues.

For example, it may be possible to use telemetry data from the farms to lessen the effect, or to replace existing radar systems with more modern units that can handle the noise. The government should invest in a research programme to find the best solution, the group concludes.

M. BETTS/GETTY

In the United States, which is selecting a new site for a high-containment livestock disease lab, a report has warned that an outbreak from such a lab could cost more than $4 billion, depending on where the lab is placed. Five sites shortlisted to replace the ageing Plum Island biosecurity lab off Long Island in New York are all on the mainland.

Phoenix digger uncovers ice in Martian soil

NASA's Phoenix lander watched some bright nuggets, exposed in a trench it had dug near the martian north pole, disappear over the course of four days last week.

That disappearance, mission officials say, means that the material must be made of ice and not the other alternative, salt. The ice sublimated away — that is, turned directly to vapour. "Salt can't do that," says Peter Smith of the University of Arizona in Tucson, principal investigator for the mission.

The lander is now digging other trenches, as well as scooping soil for chemical analysis.

International stem-cell collaborations launched

California's stem-cell institute has penned two international research agreements, including a pact with Canada that will bring in at least US$100 million for cancer-related studies.

The California Institute for Regenerative Medicine (CIRM) will form partnerships with a consortium of Canadian research facilities and the state government of Victoria in Australia, officials announced at last week's BIO International Convention in San Diego, California.

CIRM will decide this autumn precisely how much it will contribute to the partnership with the Canadian Cancer Stem Cell Consortium, which involves agencies such as the Canadian Institutes of Health Research in Ontario. Canadian officials say they have committed $100 million over three years, with the possibility of expanding to $250 million within five years.

Under the Australian agreement, CIRM-funded researchers in California would collaborate with colleagues at institutes in Victoria, such as Monash University — the former home of CIRM president Alan Trounson.

Latest satellite launches to monitor sea level

Global sea level, as it inches ever higher, has a new eye keeping watch on it from orbit.

On 20 June, the US and French space agencies launched the next-generation ocean topography satellite, the Ocean Surface Topography Mission, or Jason-2. The satellite will continue to monitor sea-surface height, a continuous series of measurements that began with the present Jason-1 mission.

Jason-2 is a joint venture of two US agencies (NASA and the National Oceanic and Atmospheric Administration), France's space agency (CNES), and the European satellite agency EUMETSAT.

Jason-1 is still operating; Jason-2 will orbit beneath it during its commissioning phase.

TUNGUSKA AT 100

The most dramatic cosmic impact in recent history has gathered up almost as many weird explanations as it knocked down trees, writes **Duncan Steel**.

Sooner or later, it was bound to happen. On June 30, 1908, Moscow escaped destruction by three hours and four thousand kilometers — a margin invisibly small by the standards of the universe.

So begins *Rendezvous with Rama*, a 1972 novel by Arthur C. Clarke in which mankind learns the hard way about the dangers posed by incoming asteroids. The 2077 impact in northern Italy that Clarke goes on to describe is fictional: the 1908 blast was real. The early morning of 30 June 1908 saw, in an area around the Stony Tunguska river, the most explosive cosmic impact in recent history, hundreds of times more powerful than the atomic weapons set off over Hiroshima and Nagasaki.

And yet, in part because it happened so far from civilization, and in part because it left no crater, it has not always been recognized as such. For decades it existed in a strange realm between science and pseudoscience, blamed on antimatter, black holes and alien spacecraft as easily as on a very fast bit of interplanetary refuse, and developing a mystique that has seen it associated with everything from energy drinks and rock bands to military missiles and *The X-Files*.

The approximate site of the blast's epicentre is now marked by a totem pole that researchers have dedicated to Agdy, the god of thunder in local mythology. Getting there is quite a trek, but the fascination of the site still draws an intermittent stream of scientists to the remote wilderness about 1,000 kilometres north of Lake Baikal; they leave offerings at the totem pole to commemorate the trek. In the years directly after the blast, though, no one came at all. The first researchers did not arrive until the 1920s.

NAT. HIST. MUS., LONDON

In 1929, more than 20 years after the explosion, dead trees at Tunguska stand witness to the scale of the event.

That does not mean there was no significant contemporary evidence to bring to bear. Siberia was and is an empty place — but a blast which, had it happened over Chicago, would have been heard from Georgia to the Dakotas, still drew a lot of attention. In the days following the blast, A. V. Voznesenskij, the director of the Irkutsk magnetic and meteorological observatory near Lake Baikal, began collecting accounts that are vivid with detail[1]. There are people being knocked off their feet, a man needing to hold onto his plough to avoid being swept away by a powerful wind, the feeling of great heat "as if my shirt had caught fire", herds of hundreds of reindeer being killed, trees set alight by the radiance of the fireball only for the flames to be snuffed out by the subsequent blast wave. And the reports are unequivocal on the source of the blast. G. K. Kulesh, head of a meteorological station at Kurensk, 200 kilometres from the epicentre, told Voznesenskij that:

> There appeared in the northwest a fiery column ... in the form of a spear. When the column disappeared, there were heard five strong, abrupt bangs, like from a cannon, following quickly and distinctly one after another ... there had been a strong shaking of the ground, such that the window glass was broken in the houses ... It is probably established that a meteorite of very enormous dimensions had fallen.

"A meteorite of very enormous dimensions had fallen."
— G. K. Kulesh

In the days after the blast, much of Europe experienced eerie 'bright nights': readers wrote to *The Times* in London, remarking that its columns could be read outdoors at midnight. Polarization measurements are consistent with this being due to sunlight scattered by dust in the very high atmosphere; observatories recorded increased atmospheric opacity and scattering across the Northern Hemisphere. This spreading dust may have been due to a plume ejected backwards along the incoming object's path by its explosion. Such plumes were seen on Jupiter when the fragments of Comet Shoemaker-Levy 9 slammed into it in 1994; hydrodynamic modelling by Mark Boslough and his colleagues at Sandia National Laboratories in Albuquerque, New Mexico, indicates that a similar terrestrial plume could be expected for an impact such as that at Tunguska[2].

There was, however, one good reason to doubt that a small asteroid was involved: the belief of the time that this would deliver a valuable hunk of iron to the surface. The Russian meteorite hunter Leonid Kulik, who led the first expedition to the epicentre in the 1920s, obtained funding from the Soviet government on the basis that he would find a valuable ore body there. But when he reached his goal in 1927 he found no metal. Nor did he find the crater that an impact was expected to leave. (There are now claims that nearby Lake Cheko might be such a crater, but these are widely disputed.) There were clear signs of violence — trees knocked flat over a vast swath of land — but no big hole in the ground. What could have happened?

In 1930, US astrophysicist Harlow Shapley suggested that the lack of a crater was due to the nature of the impactor. If it had been a comet, and comets were light and fluffy, then it would have exploded at altitude. This idea persisted for decades: in 1982 some planetary scientists were willing to postulate the extraordinarily low density of 3 kilograms per cubic metre in order to explain Tunguska in terms of the blast from a disintegrating comet.

Other explanations were even more far fetched than candyfloss comets. Soviet science-fiction author Alexander Kazantsev realized, as Shapley had, that the best explanation involved an explosion at altitude, and suggested in 1946 that a nuclear-powered alien spaceship exploding just before landing might have been the culprit, an idea taken up eagerly and earnestly in the following decades.

A more scientifically promising possibility was naturally occurring antimatter, a suggestion made independently by various people at various times. In 1940, Vladimir Rojansky of Union College, Schenectady, New York, suggested that some meteors and comets might

be made of antimatter — 'contraterrene' matter in the terms of the time — and that their odd behaviour might be detectable[3]. (More than 30 years later Rojansky suggested that it would be worth checking if Comet Kohoutek was one of the antimatter ones.) In 1941, Lincoln LaPaz of Ohio State University in Columbus published two articles in the magazine *Popular Astronomy* that argued that large terrestrial craters and the craterless Tunguska explosion were bothdue to antimatter meteors; he later wrote to the Soviet Academy of Sciences suggesting a search for anomalous isotopes at the site.

Radioactive remains

More than a decade later, Philip Wyatt, a graduate student at Florida State University in Tallahassee, and Boris Podolsky, author of a famous paper with Einstein exploring apparent paradoxes of quantum mechanics, went to a movie in which antimatter featured. Podolsky pointed Wyatt towards Rojansky's 1940 paper and suggested he look into the impacts idea. Wyatt — now the chief executive of the Wyatt Technology Corporation in Santa Barbara, California — says that he was "mostly interested in looking for residual radioactivity" and published some ideas on the subject in *Nature*[4].

This notion was expanded on by three eminent American scientists (including 1960 Nobel Prize winner Willard Libby and Clyde Cowan, co-discoverer of the neutrino) in 1965. Libby, the original developer of the carbon-14 dating technique, found support for the idea of an antimatter impact from what seemed to be an elevated carbon-14 level in tree rings around the world in 1909, suggesting that significant quantities of the isotope had been created by radiation given off when the antimatter annihilated itself on contact with the thicker layers of the atmosphere[5]. Even at the time, though, there were good arguments against the idea: among other things, the first gamma-ray-detecting satellites were not seeing the tell-tale radiation from antimatter annihilation elsewhere in the nearby cosmos.

Even more extreme, in 1973 two University

The totem of the thunder god at the Tunguska site.

M. E. BAILEY/ARMAGH OBSERVATORY

of Texas physicists suggested that the cause was a black hole passing through Earth[6]. This was nothing if not fashionable: miniature black holes had just been postulated by Stephen Hawking as after-effects of the Big Bang. Again the explanation was incomplete and its implications — an exit on the other side of the planet, and a seismic signal lasting well after the initial impact — unobserved. Similar caveats apply to the intriguing hybrid idea, aired as recently as 1989, that the culprit was a deuterium-rich comet turned into a hydrogen bomb by the heat and pressure of its arrival in the atmosphere.

Another approach has been to suggest that, despite the straightforward implications of eyewitness accounts of a bright object zipping across the sky, the source of the blast was in fact beneath the surface. A recent example is a claim that it was due to a 10-million-tonne belch of methane that subsequently exploded high in the sky. Others see a geophysical source involving peculiar tectonic behaviour.

M. E. BAILEY/ARMAGH OBSERVATORY

The Churgin creek flows past Tunguska's epicentre, its forests regrown.

Hammer time

The fact that such ideas were entertained (and still are, in some circles) speaks both of a certain fascination with the fanciful and the abiding need to explain that confusing lack of a crater. The fact that, by the 1960s, various craters around the world had been accepted as meteorite strikes meant that the anomalous lack seemed all the more confusing. In 1993 that confusion was allayed, at least for most people, by Chris Chyba, Kevin Zahnle and Paul Thomas[7]. With the help of computer simulations derived from nuclear weapons' tests they showed that a solid, stony object about 50 metres across — the most likely sort of thing in that size range to hit the Earth — would not be expected to reach the ground. There was no need to invoke weirdly low cometary densities — at the relevant speeds the shock wave generated within a solid body as it slams into the atmosphere would rip up an everyday rock just fine. Formations such as Meteor Crater in Arizona (see page 1172) are left by tougher impactors made of metal; the shock waves don't get the better of them until they've reached the ground.

A similar explanation was arrived at by Jack Hills, working at Los Alamos National Laboratory in New Mexico with Patrick Goda[8], and both teams had been to some extent pre-empted by a Soviet team led by V. P. Korobeinikov, the work of which had not been widely appreciated in the West[9]. These various models led to an estimate that the blast was equivalent to about 15 megatonnes of high explosive — bigger than all but the very largest thermonuclear weapons. However, work by Boslough indicates that the energy required to fit the observed phenomena could be rather less, around 3 to 5 megatonnes.

That analysis assumes that the impactor was a stony asteroid — but a comet is still a possibility. In 1978, L'ubor Kresák suggested the Tunguska impactor was a fragment of Comet Encke[10]. The peak of an annual intense meteor shower associated with dust from Encke occurred around 30 June 1908, but because the meteors arrived from the direction of the Sun, the shower would not have been visible to the naked eye. What the eyewitnesses said about the direction of the Tunguska projectile is consistent with that idea.

Other explanations were even more far fetched than candyfloss comets.

An analysis of many hundreds of possible pre-impact orbits for the object published in 2001, by a team that had been led by the late Paolo Farinella, indicated that an asteroidal orbit was more likely than a cometary orbit[11] — but using that paper's definitions, Comet Encke, which takes just 40 months to orbit the Sun, has an asteroidal orbit. Another line of evidence, suggested in 1977, was that a comet might explain the carbon-14 signature reported by Cowan in the 1960s; a comet in space might naturally be thoroughly irradiated[12].

The question of what the object was is not purely academic. If Tunguska was indeed a 15-megatonne event, it was rather unlikely — such things are expected only every 1,500 years or so. That calculation, though, assumes that the flux of near-Earth objects is constant over time. If the population of near-Earth objects is replenished from time to time by the break-up of a comet, then shortly after that break-up, impacts from Tunguska-sized fragments will be more likely. Earth may suffer near misses from Tunguska's dark and stealthy cousins every time it passes through Encke's dust stream — fragments too small to be easily observed, but big enough to cause quite a mess if they hit.

In *Rendezvous with Rama*, Clarke's solution to the threat of impacts was an asteroid search programme aimed at ensuring that such a catastrophe could never occur again: he called it Project Spaceguard (see page 1178). This became the name of a real-life programme, and that search continues. But 50-metre objects are too small to spot far in advance of their impact. So although another Tunguska coming out of the blue is not a likely event in any given June, it is not out of the question. ■

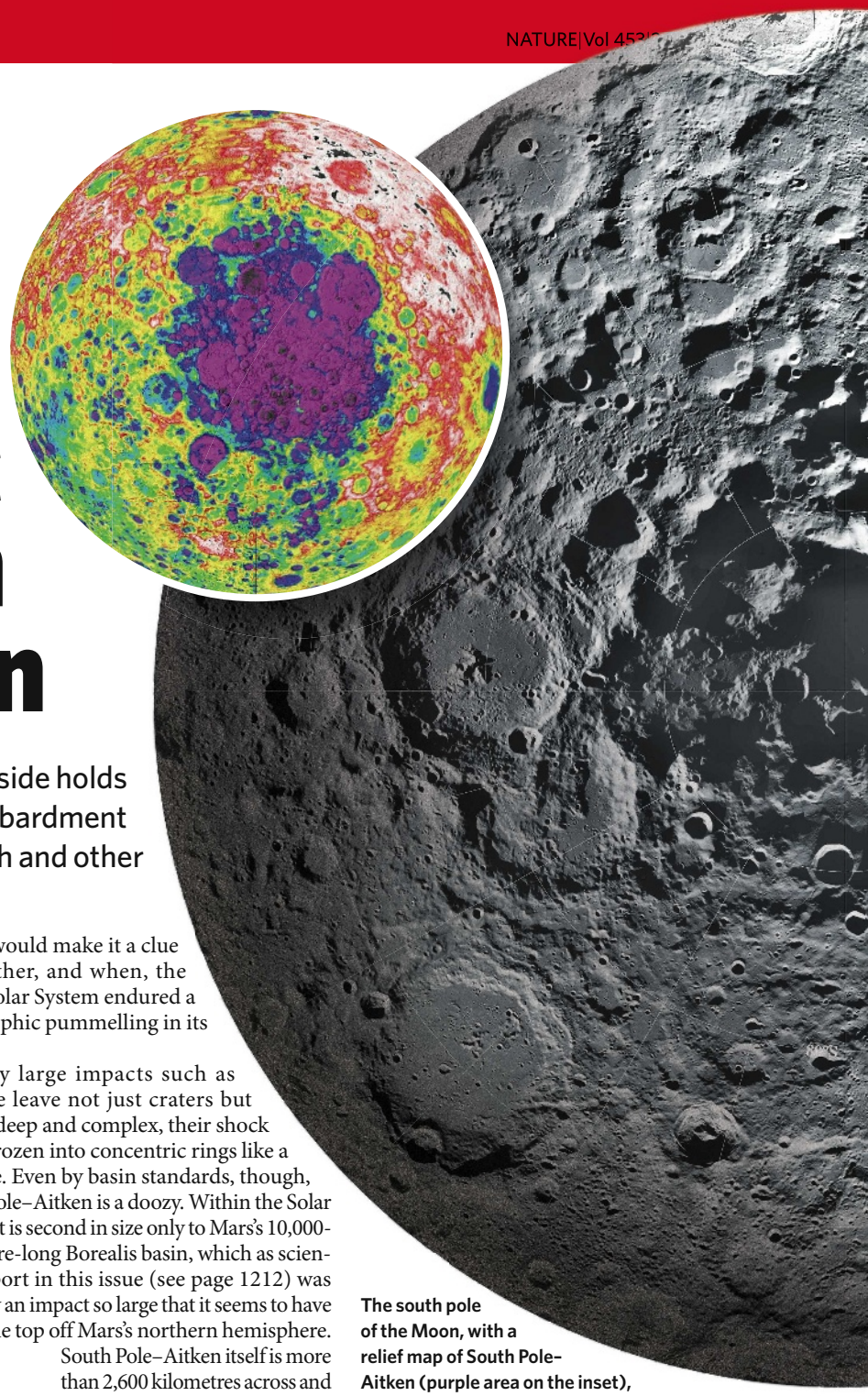
Duncan Steel is an astronomer and writer after whom Arthur C. Clarke once named a robot.

1. Longo, G. in *Comet/Asteroid Impacts and Human Society, An Interdisciplinary Approach* (eds Bobrowsky, P. T. & Rickman, H.) Ch. 18, 303–330 (Springer-Verlag, 2007).
2. Boslough, M. B. E. & Crawford, D. A. *Ann. NY Acad. Sci.* **822,** 236–282 (1997).
3. Rojansky, V. *Astrophys. J.* **91,** 257–260 (1940).
4. Wyatt, P. J. *Nature* **181,** 1194 (1958).
5. Cowan, C., Atluri, C. R. & Libby, W. F. *Nature* **206,** 861 (1965).
6. Jackson, A. A. & Ryan, M. P. *Nature* **245,** 88–89 (1973).
7. Chyba, C. F., Thomas, P. J. & Zahnle, K. J. *Nature* **361,** 40–44 (1993).
8. Hills. J. G. & Goda, M. P. *Astron. J.* **105,** 1114–1144 (1993).
9. Korobeinikov, V. P., Shurshalov, L. V. & Chushkin, P. I. *Astron. Zh+* **25,** 327–343 (1991) [In Russian].
10. Kresák, L. *Bull. Astron. Inst. Czech.* **29,** 129–134 (1978).
11. Farinella, P. *et al. Astron. Astrophys.* **377,** 1081–1097 (2001).
12. Brown, J. C. & Hughes, D. W. *Nature* **268,** 512–514 (1977).

See Editorial, page 1143, News Feature, page 1170, and Commentary, page 1178.

The hole at the bottom of the Moon

A giant crater on the lunar farside holds the key to a catastrophic bombardment that reshaped the Moon, Earth and other planets. **Eric Hand** reports.



The south pole of the Moon, with a relief map of South Pole–Aitken (purple area on the inset), mapped from the Clementine orbiter.

NASA

Dhofar 961 wasn't like the other Moon rocks. Looking at its freshly cut face, geochemist Randy Korotev noticed immediately how dark it was — almost purple — and that it contained big metallic grains. It was so different from anything he'd seen before that he began to wonder. Was it from the 'big one'?

Korotev, of Washington University in St Louis, Missouri, already knew that Dhofar 961 was a piece of the Moon, chipped off by some anonymous impact so that it escaped the Moon's feeble gravitational grasp and succumbed to Earth's. Tens of thousands of years ago, Dhofar 961 fell into the Oman desert. A few years ago, it fell into the hands of collectors eager to make a buck. After fruitless searches on eBay, Korotev found a reputable dealer online selling a 6-gram piece of Dhofar 961 for US$1,000 per gram — 30 times more expensive than gold. Korotev bought a sliver, and sacrificed a third of it for a chemical analysis that confirmed his suspicions.

> **"The Moon is a witness plate for what happened on the Earth."**
> — David Kring

Unlike the other 59 known lunar meteorites, which have chemical compositions that trace back to three specific regions on the Moon, Dhofar 961 probably hails from a fourth: a deep, dark hole at the bottom of the lunar backside, known as the South Pole–Aitken basin. It marks the site of the biggest-known blast the Moon has seen. And trapped within Dhofar 961 might be a record of that event, which would make it a clue to whether, and when, the inner Solar System endured a catastrophic pummelling in its youth.

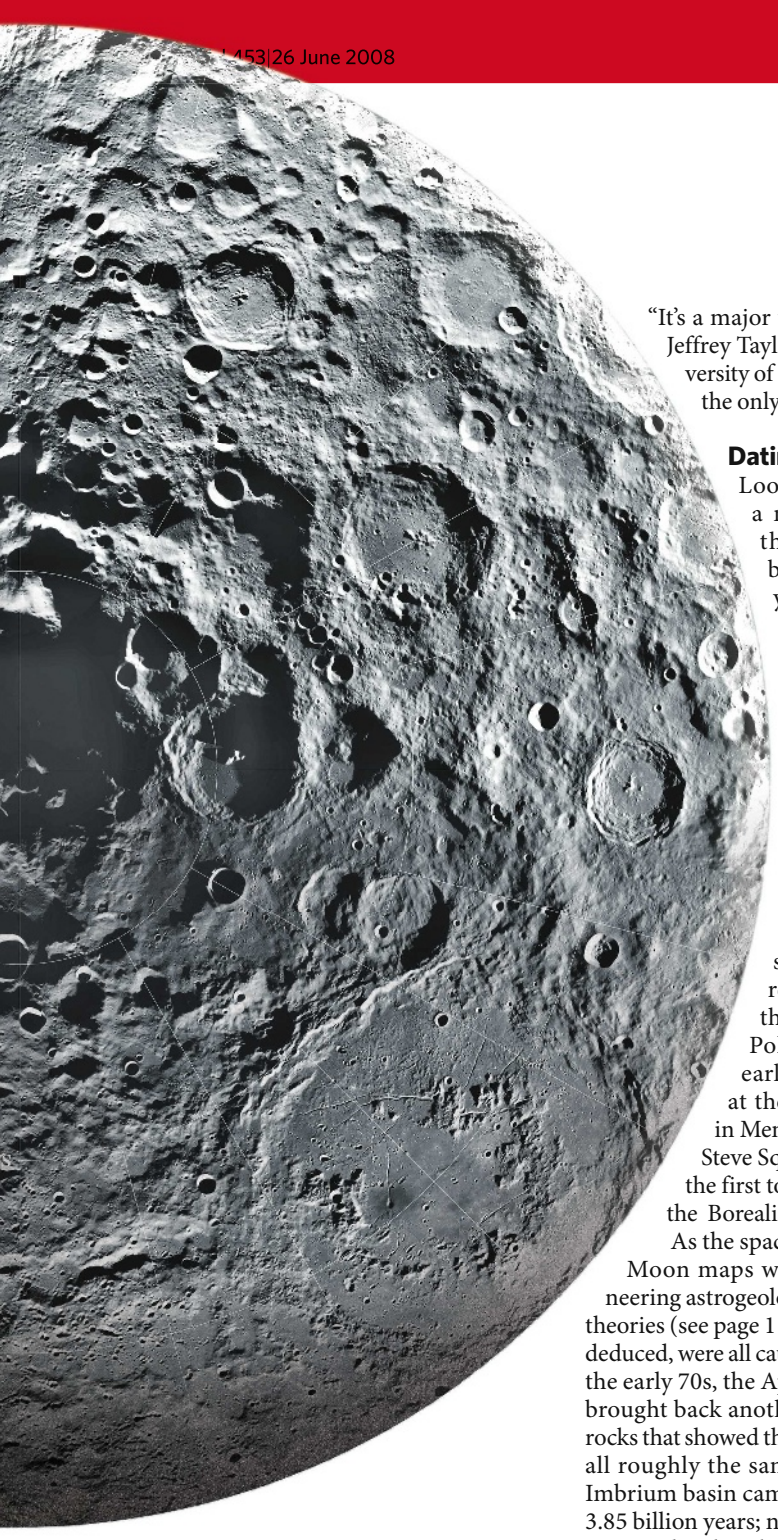
Really large impacts such as this one leave not just craters but basins, deep and complex, their shock waves frozen into concentric rings like a bullseye. Even by basin standards, though, South Pole–Aitken is a doozy. Within the Solar System it is second in size only to Mars's 10,000-kilometre-long Borealis basin, which as scientists report in this issue (see page 1212) was made by an impact so large that it seems to have sliced the top off Mars's northern hemisphere. South Pole–Aitken itself is more than 2,600 kilometres across and 12 kilometres deep, big enough to blot out half of China and hide the highest mountains of Tibet.

South Pole–Aitken is not only the biggest basin on the Moon, but also the oldest, based on the relative chronology that geologists piece together by mapping the way craters overlap each other. An absolute date for it is, however, unknown.

Just after the Solar System formed 4.6 billion years ago, leftover planetesimals regularly blasted the newborn planets. The barrage even knocked off enough of Earth to create the Moon in the first place. By 3.8 billion years ago, impact rates had tailed off to a level not too different from those of today (see graphic).

The question is what happened in between: did the impacts decrease smoothly, or was there, as many scientists suspect, a big spike 3.9 billion years ago? Given South Pole–Aitken's prominence at the bottom of the cratering heap, its age provides a crucial constraint on this 'late heavy bombardment' or 'lunar cataclysm'. An early date for South Pole–Aitken means a broader peak in the bombardment rate, or possibly a steady rate throughout the period. A later date speaks to cataclysm.

This is why the US National Academies last year called dating South Pole–Aitken the most important goal in lunar science. Date the basin, and you test the idea of a cataclysm, with



P. SPUDIS/LUNAR & PLANETARY INST.

all its ensuing implications. It tells you what was happening on the Moon early on and, by inference, what was going on in the rest of the Solar System. It tells you whether the inner planets got smacked suddenly in an atmosphere-annihilating blast of impacts. And that has implications for the origins of life. Was the great bombardment so severe that it sterilized any life that had got started before then? Did it create the hellish conditions that many of the earliest life-forms seem to have endured? Could it even have moved life from one planet to the next, throwing travellers such as Dhofar 961 from surface to surface, complete with bacterial hitchhikers?

"It's a major unsolved problem," says Jeffrey Taylor, a geologist at the University of Hawaii. "And the Moon is the only place we can address it."

Dating the impacts

Look at the Moon for even a moment, and it's clear that the place has been brutalized. Yet for many years, scientists thought that its cratered surface resulted from inner turmoil rather than outer. Impact craters were mistakenly identified as volcanic calderas, the remnants of explosive eruptions. "You have no idea how pervasive this idea was — that volcanics were responsible for everything on the Moon," says Don Wilhelms, a retired geologist who was the first to map the South Pole–Aitken basin in the early 1970s, when working at the US Geological Survey in Menlo Park, California. With Steve Squyres, Wilhelms was also the first to propose the existence of the Borealis basin on Mars[1].

As the space race took off, rich new Moon maps were produced, and pioneering astrogeologists buried the volcanic theories (see page 1164). The big basins, they deduced, were all caused by impacts. Then, by the early 70s, the Apollo astronauts had brought back another surprise: Moon rocks that showed that the impacts were all roughly the same age. The huge Imbrium basin came in at an age of 3.85 billion years; nearby Nectaris, separated in the relative chronology by hundreds of substantial craters, was just 50 million years younger. Nothing was older than 4 billion years.

In 1973, Fouad Tera and his colleagues at the California Institute of Technology in Pasadena first used the term 'cataclysm' to explain the extreme pace of the impacts. "It must in any event have been quite a show from the Earth, assuming you had a really good bunker to watch from," they wrote in an abstract to that year's Lunar and Planetary Science Conference.

It wasn't just a 'show from the Earth', though; it was the greatest show the Earth itself has ever experienced — the sort of show you're lucky to come through intact. The Moon and Earth are so close to each other that whatever happened on the Moon also happened on Earth — and then some. The record has been lost on Earth because most impact craters are erased through weathering, erosion and the continuous churn of plate tectonics. That makes the Moon "a witness plate for what happened on the Earth," says David Kring, a geologist at the Lunar and Planetary Institute in Houston, Texas.

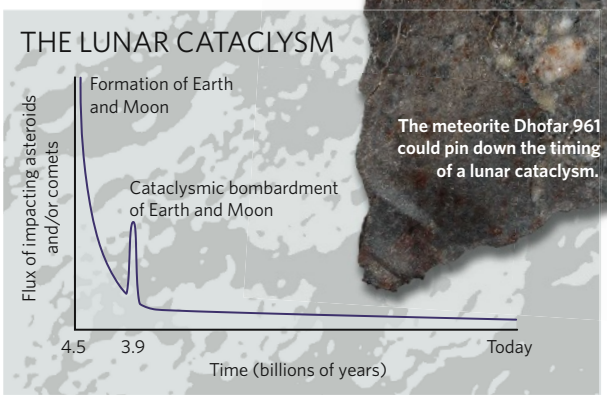
And what a bad time it was. To model what Earth went through, Kring scaled up what happened to the Moon by a factor of 13 to account for the fact that Earth is a much larger target[2]. Given Wilhelms' estimate of 15 major lunar-impact basins in the 50 million years between Nectaris and Imbrium, this meant to Kring's team that a projectile big enough to form a 20-kilometre crater hit Earth every few thousand years. Every million years, something would come along big enough to make a 1,000-kilometre basin. Such impacts would have vaporized Earth's oceans and steam-sterilized the surface; Kring says an atmosphere of rock vapour could linger for thousands of years after the impact.

Here's the crazy part: Kring's estimate is, in fact, very conservative. Earth's strong gravity could attract impactors at a frequency as much as 500 times higher than the Moon would. Moreover, Kring does not include in his calculations the 30 other huge basins that, according to Wilhelms, were formed after South Pole–Aitken and before Nectaris. If South Pole–Aitken turns out not to be significantly older than Nectaris, then the frequency of doomsday rocks hitting Earth rises yet higher.

What's more, Wilhelms' basin count — a baseline for many studies — is now old,

R. KOROTEV

The meteorite Dhofar 961 could pin down the timing of a lunar cataclysm.



and is conservative itself. Herbert Frey, at the Goddard Space Flight Center in Greenbelt, Maryland, recently finished a new hunt for basins, based on topographical data collected by the Clementine Moon orbiter. At the 2008 Lunar and Planetary Science Conference, Frey's team reported 92 basins bigger than 300 kilometres across — twice as many as Wilhelms. "We've grossly underestimated the actual flux of objects that hit the Moon," Frey says. "It means that the Earth was probably not a very good place to be 4 billion years ago."

"It means that the Earth was probably not a very good place to be 4 billion years ago." — Herbert Frey

Yet some astrobiologists say that a cataclysm may have catalysed the origin of life rather than snuffed it out. A stream of comets or asteroids hitting the planet would have brought foreign organic material to Earth. The bombardment might have pierced the crust, stirring up deep convective currents in the mantle in such a way as to establish early continental crust. And, although surface oceans might have been stripped away, subsurface water and heat could have nourished heat-loving organisms. It's probably no coincidence that in phylogenetic trees of life, the roots of the three major branches — bacteria, archaea and eukaryotes — tend to be heat-loving[3].

Ultimate causes

One thing the lunar rocks make clear is that the bombardment dropped off pretty quickly about 3.85 billion years ago. But what was doing the pummelling in the first place? Some have claimed it was simply the expected collisions of things left over from the formation of the Solar System, but there's no obvious way there would have been enough projectiles at the beginning to last as long as would be needed for that. Others say the cataclysm was a fresh spike of new bombardments, but what would have caused such an influx of new impactors? "Some people didn't like it because they couldn't think of a mechanism," says Wilhelms.

A few years ago, one possible explanation surfaced from researchers based in Brazil, the United States and in Nice, France. The team developed a dynamical model for the Solar System that explained why Uranus and Neptune circle the Sun farther out and more eccentrically than expected[4]. In their model (sometimes called the Nice model), they started the infant Solar System with Neptune's orbit inside that of Uranus, and let the clock run. Some 700 million years later, Jupiter and Saturn fell into an orbital pattern, and the resulting gravitational pull caused Uranus and Neptune to be kicked farther away from the Sun. That in turn disrupted a massive disk of icy comets in the Kuiper belt beyond Pluto, and sent them hurtling into the inner Solar System.

There's one problem with all this. The chemical composition of material within lunar craters, as well as their size distribution, matches nicely with asteroids, not comets — suggesting that asteroids were the main, or most recent, impactors during the bombardment. The Nice modellers have an answer for that: the changes in Jupiter's and Saturn's orbits may have also disrupted the asteroid belt between Mars and Jupiter. That could have been enough to send asteroids smashing into Earth, the Moon and more.

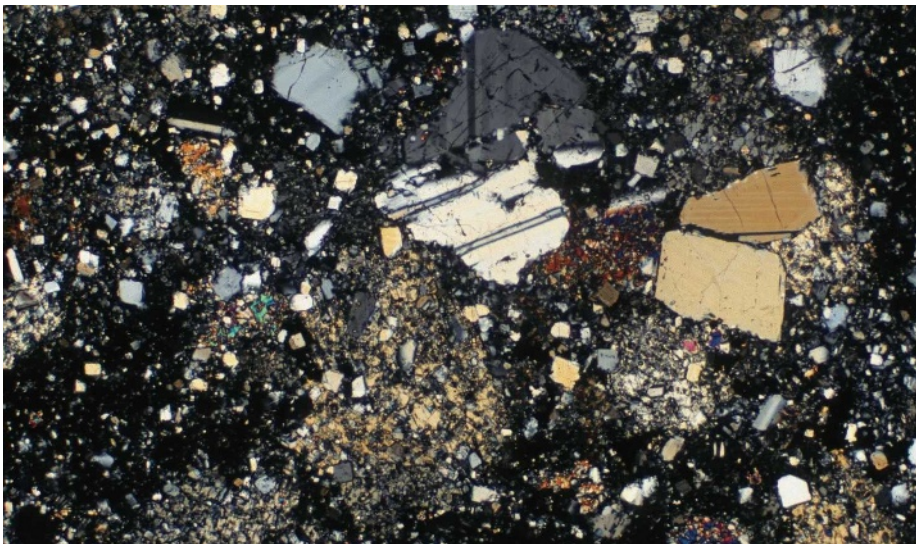
Korotev says he is now a believer in the lunar cataclysm, thanks in part to the Nice model. Other work is resolving his other long-standing problem with the cataclysm hypothesis: his belief that the dating of most Moon rocks to around 3.9 billion years ago is the result of an artificial selection bias. He and some other researchers have argued that the astronauts might have simply kept picking up rocks from the Imbrium impact over and over again, and that scientists interpreted them as being from different impacts.

Astronauts like John Young (shown here on the Apollo 16 mission in 1972) had geological training to choose the most promising rocks for analysis.

Marc Norman, a cosmochemist at the Australian National University in Canberra, continues to wring new dates from the Apollo collection that may counter this challenge. Norman has been looking at impact melts, the recrystallized remains that contain an isotopic record of a rock being melted by an impact. In one Apollo rock, Norman dated 21 impact melts to within a 200-million-year window[5]. And he found that the melts fell into a number of age clusters, which he interprets as representing four different impact events. If his interpretation is correct, that would be more evidence that multiple big basins were formed within the narrow time period of the bombardment.

Other evidence is coming from meteorites, which unlike the geographically constrained Apollo rock collection are thought to have been hacked from all over the Moon. Working with Kring, Barbara Cohen, now at the Marshall Space Flight Center in Huntsville, Alabama, analysed four meteorites containing impact melts representing seven to nine impact events. None of the melts, she found,

R. ZEIGLER



Time capsule: impact-melt Moon rocks, such as the slice shown here, all date from around the same time.

NASA

was older than 3.92 billion years[6]. Even the notorious martian meteorite Allan Hills 84001 — famous a decade ago for claims that it contained evidence of life — offers support for the lunar cataclysm theory: parts of the 4.5-billion-year-old rock were altered in some sort of major event 3.9 billion years ago.

Reaching for the Moon

Still, some lunar scientists are not satisfied with Moon rocks fetched by astronauts or fallen from the sky. The best way to date South Pole–Aitken, they say, is to go there, get a rock, and date it.

Sending a simple robotic lander would be relatively cheap, but the robot's dating capabilities would not be good enough. Radioisotope dating requires a mass spectrometer, and one small enough to fly on a lander would have uncertainties of 10%. On a 4-billion-year-old rock, that's 400 million years — exactly the sort of error that a mission travelling to date South Pole–Aitken is supposed to dispose of, not create. "You want to be able to say that your 4.2-billion-year-old age is different from a 4.0," explains Taylor. "It requires more accuracy than we have at present."

And so a group of lunar scientists is pushing for a South Pole–Aitken sample return mission, which would be the first lunar sample return since the last Soviet Luna spacecraft returned 170 grams of soil in 1976. The group, led by Brad Jolliff of Washington University in St Louis, plan to propose a mission for the next NASA New Frontiers competition, a mission class capped at $650 million. NASA intends to start accepting proposals in December, with eventual selection in 2010 and a launch date no earlier than 2015.

In the most recent New Frontiers contest in 2005, a South Pole–Aitken mission called Moonrise made it to the final round but was ultimately bested by Juno, which is set to launch towards Jupiter in 2011. At the time, the risk of doing a sample return mission, with its many stages and components, was considered riskier than a simple orbiter such as Juno, says Jolliff, who was the deputy principal investigator on that proposal.

"Landing in the middle of a field and getting a scoop of dirt is not going to give you the answer you need."
— Marc Norman

The old Moonrise project proposed two separate landers: one to go near the basin rim and the other, the centre, where impact melts are apt to be concentrated. Jolliff is leaving open the option to send just one lander to the basin centre. By 2020, if NASA's plans to return people to the Moon are realized, astronauts could already be encamped at the nearby Shackleton crater, placing them near the rim of the South Pole–Aitken basin. The Moonrise lander would collect rock and soil, and return to Earth with about a kilogram of material, Jolliff says.

Samples from a revamped Moonrise mission would allow lunar scientists to date many impact-melt crystals. The oldest, and most frequent, dates should correspond to the South Pole–Aitken impact. But other impact-melt dates would undoubtedly pollute the picture, as there are half a dozen other large basins within South Pole–Aitken. And debate would continue. "Landing in the middle of a field and getting a scoop of dirt is not going to give you the answer you need," says Norman.

In fact, Norman advocates both robotic and manned missions to the Moon, saying both are needed for a balanced exploration programme. "My feeling about sample return may be a little more nuanced than simply humans versus robots," he says. "Both can do the job provided we do the geologic homework, and neither will do an adequate job if we don't." Others, however, argue that the mystery of the great backside basin won't be solved until a human goes to the source and plucks a rock from within its blast shadow.

But returning people to the Moon will cost at least $230 billion over two decades (according to the US Government Accountability Office), compared with the New Frontiers $650-million cut-off. And Korotev thinks he can solve some of the mystery simply by dating the half-gram piece of the Moon he bought for $542 (plus $12 shipping and insurance). Later this year, he will share his precious sliver of South Pole–Aitken with Cohen. In Huntsville, she plans to sink a diamond-tipped drill bit into Dhofar 961 and extract several cores, each as fine as a human hair.

After vaporizing the samples with a laser, she will measure argon gas that has been trapped inside the rock crystal lattice for billions of years. Counting those atoms might allow her to count back in time to the blistering crucible of the bombardment. If she's successful, she will extract a date desired by lunar scientists for many moons — a big message from a little bottle. ■

Eric Hand covers physical sciences for *Nature* from the Washington DC office.

1. Wilhelms, D. E. & Squyres, S. W. *Nature* **309,** 138–140 (1989).
2. Kring, D. & Cohen, B. *J. Geophys. Res.* **107,** 41–46 (2002).
3. Zahnle, K. & Sleep, N. H. in *Comets and the Origin of Life* (eds Thomas, P., Chyba, C. & McKay, C., eds) 175–208 (Springer, 1997).
4. Gomes, R. *et al. Nature* **435,** 466–469 (2005).
5. Norman, M. D., Duncan, R. A. & Huard J. J. *Geochim. Cosmochim. Acta* **70,** 6032–6049 (2006).
6. Cohen, B. A., Swindle, T. D. & Kring, D. A. *et al. Science* **290,** 1754–1756 (2000).

See Editorial, page 1143, and News Feature, page 1164.

THE BURGER BAR THAT SAVED THE WORLD

Fewer people are searching for near-Earth asteroids, astronomer David Morrison said in the 1990s, than work a shift in a small McDonalds. But that group — a little larger now — has over the past two decades discovered a host of happily harmless rocks, and in doing so reduced the risk of an unknown asteroid blighting civilization (see page 1178). **David Chandler** puts together the story in the words of those who watched, and those who watched the watchers.

Clark Chapman: About 60 years ago, there were some prescient things written by Ernst Öpik, by Ralph Baldwin, and by Fletcher Watson. Only a handful of near-Earth asteroids had been discovered, but they came up with order-of-magnitude-correct understandings about how often a bad thing would happen.

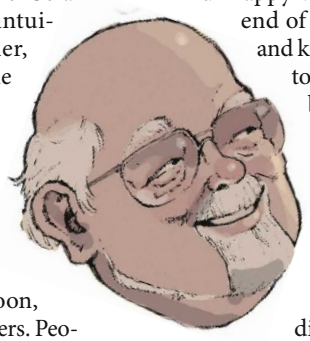
David Morrison: That understanding arose almost without reference to Tunguska. Öpik and Gene Shoemaker did some kind-of-heroic calculations based on almost zero data. In the 1950s, we only knew of a few Earth-crossing asteroids and had data on a couple of comets that had come into the inner Solar System. And they, using physical intuition and consistent with each other, made the first predictions of what the impact flux might be. Before that, it was pure arm-waving. Öpik and Shoemaker quantified it, and within the right order of magnitude.

Carolyn Shoemaker: When the first Apollo mapping studies were done, trying to get better photos of the Moon, they became much more aware of craters. People like Gene were convinced that the majority were caused by impacts. They looked so much like those on the Earth. Not only Meteor Crater, but also nuclear-bomb craters. So he was convinced they were caused by impacts.

Rusty Schweickart: I tramped all around Meteor Crater with him, as did all of us who were part of the Apollo programme back then. Nobody knew whether features on the Moon were impact or volcanic. So we went all over the world to volcanic sites, to impact sites, with the top people in the world, including Gene. Frankly, very few people thought the main cause was volcanic. People did think there might be some volcanic features, which of course there are.

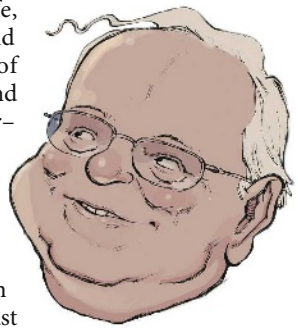
1978
Known near-Earth asteroids, end June
Large = 41
Total = 48

Chapman: Gene got very unhappy with NASA, there at the end of the Apollo programme, and kind of consciously wanted to get away from it all. He'd been interested in craters on the Earth, and clearly was aware that they were related to asteroids in the sky, and got interested in going off to sit in observatories on lonely mountain tops to discover them in the early 70s. He hooked up with Glo Helin to do a joint observing programme.



Steve Ostro: Observations of near-Earth objects were growing — the stuff by Helin and by Shoemaker — and we started getting some radar opportunities on newly discovered objects. In 1979, I met all the asteroid people and got very enthused, and then I was pretty sold on asteroids and wrote my own observing proposals. Within a year I was basically doing that more than anything else.

Brian Marsden: Helin and Shoemaker fell out and had their own separate programmes. It's always been a bit of a mystery — I've never fully understood it. Gene decided he could get Carolyn, his wife, involved. Carolyn did all the examining of the films, so she found all those Shoemaker–Levy comets. Gene was more the big picture: he did do a fair bit of the actual drudge work you might say, but Carolyn did a lot more of it, just as Helin and her people had done.



Shoemaker: I think there were only nine near-Earth asteroids known when Gene started the programme with Helin. They found a few, not

ASTEROID NUMBERS FROM A. CHAMBERLIN, JPL

ILLUSTRATIONS BY D. PARKINS

Who's who and what's what

Clark Chapman, of the Southwest Research Institute in Boulder, Colorado, has studied asteroids and comets throughout his career. In the early 1980s he became interested in the risks posed by **near-Earth asteroids**, which are those that come within 45 million kilometres or so of the Earth's orbit. *Cosmic Catastrophes*, by Chapman and **David Morrison**, an astronomer at NASA's Ames Research Center in Moffett Field, California, drew attention to the risk in the late 1980s. ● **Tunguska** was the Earth's most destructive twentieth-century impact (see page 1157). ● The late **Gene Shoemaker** of the US Geological Survey in Flagstaff, Arizona, went from studying craters to discovering asteroids with his wife **Carolyn**. At **Meteor Crater** in Arizona (see page 1172) he trained astronauts such as **Rusty Schweickart**, who flew on Apollo 9. ● **Eleanor 'Glo' Helin** discovered or co-discovered 872 asteroids. Before she retired, she worked at NASA's Jet Propulsion Laboratory (JPL) in Pasadena, California, as does **Steven Ostro**. ● **Brian Marsden** ran the

many, and Gene's feeling was that we have to find them a lot faster — develop another programme to find more and faster. We went to using film that had a very fast emulsion, and we'd take two films, 45 minutes apart, so anything near would seem to move.

Then Tom Gehrels started Spacewatch in the 1980s, which Jim Scotti continued.

Chapman: Now Tom, he was sort of 'Mr Asteroid' in the 1960s. He'd been studying asteroids just for the sake of their scientific interest — measuring their light curves and other physical properties.

1988
$N_{Large} = 81$
$N_{All} = 108$

Tom Gehrels: We had a CCD on a telescope, the dedicated telescope that was the Spacewatch telescope, by 1983. Soviet submarines had already used CCDs in the 1970s, and our colleagues there were quite free to talk about it, which was quite remarkable, so I learned a lot about CCDs. The first asteroids were discovered by Spacewatch in 1984, and the first near-Earth asteroids were found with that system in about 1986. This was done against everybody's opposition, including Shoemaker's — he fought like a tomcat.

Shoemaker: There was some communication [with Spacewatch], and I guess competition in a sense. They were able to see much fainter magnitudes than we could. That's been their big contribution.

Killing the dinosaurs

Gehrels: Öpik and Harold Urey had warned very clearly of the impact hazard, but I had totally ignored that, and it was the Alvarez paper that really woke that up.

Morrison: Until the Alvarez paper on the K/T boundary there was no reason to think that any impact, short of some humongous thing, would have any global consequences. The truly remarkable thing was that such a small impact — that has no effect on the Earth's rotation, its axis, its magnetic field, completely negligible physically — could nevertheless wipe out most of the life on the Earth. To say that the biosphere was so fragile was a real revolution. In terms of the hazard, it was key.

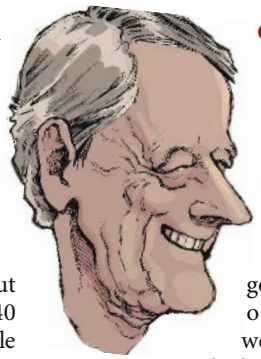
Chapman: So Bill Brunk and Shoemaker put together this group of I'm going to guess 40 people that included a few people from the military, asteroid scientists like myself, people who knew about orbits such as George Wetherill of the Carnegie Institution in Washington, and some NASA mission-planning-type people. It was just a wide-ranging discussion — it included the nature of the threat, the possible damage to the environment.

Morrison: It took a while for other people to accept the idea of an impact. Especially palaeontologists, who thought this ridiculous egotistical physicist was trying to tell them what killed the dinosaurs.

The astronomers embraced the Alvarez result very quickly, and geologists such as Shoemaker started drawing conclusions from it, but the palaeontologists resisted it equally dramatically, for quite a while.

Chapman: People were aware of Project Icarus, and there was even discussion of the politics of it. Wetherill in particular was very concerned that the project would open up the door to the use of nuclear weapons in space. All those kinds of issues were discussed during this 3- or 4-day meeting at Snowmass in Colorado. And a report was written, a very lengthy report, but it was never published.

Morrison: That group actually made the recommendation that NASA should consider how to discover and how to defend against incoming objects. It was quite early. Now, that doesn't mean it was accepted. But the guy who really was in the intellectual leadership on this was Shoemaker.



Gehrels: I was pretty good friends with Shoemaker. He didn't believe in it. He would say 'Tom, this is not for real'. He organized the meeting, and for a day and night we were not supposed to go out and see the snow or anything like that, we just worked, worked, worked, so we could put a book together. And by the time it was put together, Shoemaker was totally convinced that it was not for real. And so we never got the book off his desk until somebody in Flagstaff actually pinched it, and then copies started floating around, but the book was never published

Congress and comets

Morrison: None of the Spaceguard stuff would have happened at NASA if Congress hadn't called for it. Congress asked NASA to organize two workshops, one on detection and one on defence against asteroids. I chaired what became known as the Spaceguard workshop on detection.

Chapman: I was asked to run the first major scientific conference on near-Earth asteroids in '91. It was a pretty large meeting, 80 people or so, and it also received some national press: it was covered by *Time* magazine, I was interviewed by National Public Radio. Morrison's Spaceguard committee held one of its meetings in conjunction with the scientific conference. And then the next year was the meeting at Los Alamos on deflection, also run by NASA but with Edward Teller and Lowell Wood and others playing a prominent part.

Morrison: At Los Alamos, the astronomers, led by Shoemaker, went face to face with the weapons people led by Teller and Wood, and that was a real wake-up call for us, that this whole other world existed, which didn't speak our language, which didn't operate the way we operated.

International Astronomical Union's Minor Planet Center in Cambridge, Massachusetts, a central clearing-house for information about new asteroid and comet discoveries, until his retirement in 2006. ● In 1980, **Tom Gehrels** and Robert McMillan of the University of Arizona in Tucson started the Spacewatch asteroid-search programme, which pioneered the use of **CCDs** (charge-coupled devices) as an alternative to photographic films. ● The **Alvarez paper**, published in 1980, was a study by geologist Walter Alvarez, his Nobel-prize winning father Luis, and their colleagues. The paper established that there was a layer of iridium present all around the world at the boundary between the Cretaceous and Tertiary periods (the **K/T boundary**) 65 million years ago. This was taken as evidence of a large asteroid or comet impact that spread dust around the Earth, cutting off sunlight, cooling the climate and triggering a mass extinction event that claimed the dinosaurs, and many other types of life, among its victims. ● **Bill Brunk** was head of NASA's Planetary Astronomy programme in the early 1980s. ● **Project Icarus** was a study done by students at the Massachusetts Institute of Technology (MIT) in the 1960s that looked at the possibility of deflecting an asteroid likely to hit the Earth. ● **Los**

Shoemaker: Teller, being a man of the background he had, was interested in possibly sending something out to blow up a near-Earth object. So that became an argument.

Morrison: The cultures of the two groups could not have been more different. Just seeing the confrontation between Teller and Shoemaker was absolutely one of the memorable things in my life. Because Teller was idealized and feared by all these people. All the weapons people seemed to be beholden to him, probably for their jobs, they almost worshipped him, they fawned over him, it was always "Dr Teller". If he so much as cleared his throat everybody in the room stopped to let him have his say. In contrast, here was good old Shoemaker — nobody would ever call him Dr Shoemaker — and we'd go out drinking with him and we had a much more egalitarian sense of things. And we did open peer-reviewed publications and the weapons guys didn't. It was really a clash of two cultures.

The weapons people seemed to think the problem was an order of magnitude greater than the astronomers did. They were for rockets on the pad with nuclear bombs, practically what you would use for a ballistic missile, and shoot the thing down right before it came into the atmosphere. It took a long time for them to even grasp the concept that if you carried out a survey you could make the discovery long in advance, and that completely changes the calculation.

At one point Teller actually stood up and said that the asteroid threat, which was real, was the appropriate justification for building much bigger nuclear bombs — a hundred or a thousand times bigger than we had.

Gehrels: Shoemaker changed with Shoemaker–Levy 9. After the impact with Jupiter, he was totally converted and he really threw his weight behind that.

Shoemaker: Gene came home from a conference that was right after it had happened, I think it was in Seattle, and he said "at last, my geologist friends believe that impacts occur". Before, people would say yeah, maybe that's so, but they just didn't really have a deep conviction that impacts occurred. After Shoemaker–Levy 9, when one could see what was happening on Jupiter, yes, we knew for sure. And geologists generally, I think, came to accept impacts. Not everyone yet, but most of them. It's really kind of amazing — it's something that was easier for astronomers to accept I think than for geologists.

1998
$N_{Large} = 211$
$N_{All} = 528$

Surveys and scares

Marsden: The discovery rate has steadily gone up as the CCD surveys came along: Spacewatch in the '80s; Helin's NEAT at end of '95; then LINEAR. In '98 they really boosted up, and they were the leader.

Shoemaker: The biggest and most successful sky survey is LINEAR, which started out using one of the telescopes that belongs to the Air Force, and that developed as a combination of MIT and Air Force technology. So they were way ahead to begin with on software sorts of things — they have kind of just thrust us all off our feet. They're the sort of thing that Gene knew we had to have if we were really going to find these bodies in anything like reasonable time. We just weren't doing that with our survey, with film and the old telescope.

Marsden: When we had the 1997 XF11 situation, people had been searching for some time and finding things, but nobody had really been doing anything from the point of view of whether these objects can be a danger. For XF11 there was a possibility of its coming very close in 2028, it could have come as close as 32,000 kilometres. Because we'd had it under observation for only 3 months, it would have been impossible to say that it might not hit us a few years after that approach. There were several opportunities on subsequent approaches, particularly in 2040.

Chapman: It's important that astronomers don't appear to be Chicken Little, and lose credibility. Public assertions by supposedly credible astronomers that there was a "small" chance that the Earth would be hit in the year 2028 by the 1.6-kilometre-wide asteroid, 1997 XF11, were corrected a day later.

Andrea Milani: I went to my office and opened my e-mail, and my folder was full of mail about 1997 XF11. I was immediately quite upset. None of the scientists involved in discussing the issue actually knew how the computation of whether an asteroid could strike the Earth could be done.

There was a real lack of knowledge. It would have been better if they just said "we don't know", rather than saying something wrong. The actual conclusion in doing the post-mortem is that the two groups of scientists who were fighting about it were both wrong. Neither had the correct algorithm.

Marsden: I did in a way stir things up a little bit at that time. At the time, it was possible it really could have hit us in 2040, based on the information we had. As we got further information, that possibility went away. And that really did get people thinking.

Milani: We needed a fully nonlinear theory of impact prediction. We found a method capable of doing this and in April '99 we were ready with a paper announcing a possibility of impact for the asteroid 1999 AN10 with a probability of 10^{-9}. Our result, from our point of view, was very good precisely because we had detected a very minute possibility. From the point of view of the journals, the result was not important because the probability was minute. But of course what

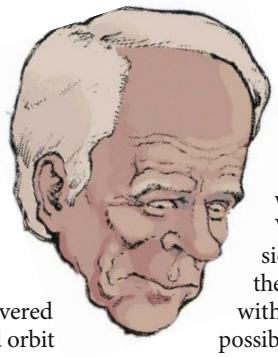
Alamos National Laboratory in New Mexico is one of the three US national laboratories concerned with nuclear-weapons design. **Edward Teller**, who was partly responsible for the design of the hydrogen bomb, and his protégé **Lowell Wood**, a nuclear-weapons designer, came from one of the other weapons labs, Lawrence Livermore National Laboratory in California, where both of them worked at the time on 'Star Wars' missile defence systems.

● **Shoemaker–Levy 9** was a comet discovered by Carolyn Shoemaker, her colleague David Levy and her husband. Pulled apart into a 'string of pearls' by a close approach to Jupiter, the comet crashed into the planet in 1994.

● After working with Gene Shoemaker, Helin went on to found the Near-Earth Asteroid Tracking (**NEAT**) survey which, like Spacewatch, used CCDs and computerized its search procedures. The Lincoln Near-Earth Asteroid Research (**LINEAR**) programme, run by NASA, the US Air Force and MIT's Lincoln Laboratory, uses similar but more advanced technology on a telescope in White Sands, New Mexico, developed for tracking satellites. It was the most powerful asteroid survey system until the advent of the Catalina Sky Survey, headed by Steve Larson of the University of Arizona, which uses two telescopes in Arizona and one in Australia. ● **Andrea Milani**

mattered was the method; the specific announcement was irrelevant.

I think it was the most important paper I ever wrote, but for a long time I was unable to publish it. Someone found this paper on our website, and I was more or less simultaneously accused of hiding this, and of spreading alarm.



Chapman: In 2004, an object was discovered by LINEAR and the nominal calculated orbit had the asteroid hitting the Earth the next day. A whole lot of people went into a real frenzy trying to get more observations of this object, because — well, if it's going to hit the Earth tomorrow, it's not a very big object, but it was about half the size of what caused Tunguska.

Europeans tried to look for it, but the weather was bad there, it was cloudy over most of the United States, and so people could not follow up on it immediately, despite considerable attempts to do so. The people at JPL, who are as expert as any, actually concluded that there was something of the order of a 30% chance that this object would hit the Earth during the next three days. Now, it turns out that the thing was much farther away and much bigger and came nowhere near the Earth. But based on the observations that were available up to that time it was a very significant probability. So we were debating late into that night, at what point should we go public with this?

That was the first incident that really raised the questions of who makes the decisions and how does the information get communicated, because before that event we've always thought of these near-Earth asteroid issues as being long-term ones, where something may be discovered now but it's going to hit in 30 years time. The idea that you'd really need to talk to people within hours just really hadn't occurred to us. Because of this event, we now realize that although the chances of something hitting tomorrow are very, very low, the way the observations are collected can certainly make it seem possible that something's going to hit tomorrow.

Schweickart: The B612 Foundation has formed a committee, working with the United Nations, to work on things such as warnings and all-clears. We're working on a decision programme, similar to the mission rules we came up with for space flight — every possible eventuality is taken into account, so when something happens, you don't start arguing about it, you just do it. We will make recommendations to the UN in September, and it will begin debating it next year. We have to set in place a process. If somebody is going to deflect something, how do you determine who does it?

Ostro: Sooner or later we will want to, or we will have to, send spacecraft to near-Earth asteroids, and come very close to them, possibly land on them. But it's very difficult to navigate around these objects, because of the unusual shapes, the spin states, and the low masses. The dynamics around a near-Earth asteroid are very different from around a big massive sphere such as the Earth — the orbits are gorgeous, geometrically intricate and complex, and every asteroid has a different dynamic environment.

2008
$N_{Large} = 743$
$N_{All} = 5,448$

Schweickart: We've hired the JPL to do a full-blown analysis of the gravity tractor which will point the way to a demonstration mission, to learn what the control parameters are.

Ostro: People have workshops on what do we do, should we deflect or blow it up, but they almost never use a realistic model of a near-Earth asteroid, they always assume a sphere or that we will know what the physical properties are before we have to do something. It's almost impossible to figure out what to do unless you know something about the object. That's where radar comes in. Right now, we've gotten radar echos from 340 asteroids.

Looking ahead

Marsden: A UK government task force produced a pretty nice report suggesting we should extend the searches down to objects 300–400 metres across. Then NASA did further studies and said we've got to go down to 140 metres and find them in 20 years, 90% of them. That's tens of thousands of objects! Over 100,000, I think. That's a big survey.

Chapman: At some level, it's going to happen anyway. At least the Pan-STARRS telescope is going to become operational any month now — it saw first light some time last year. The people running it have already spent energy on what to do if they find near-Earth asteroids on their CCDs.

People are working on the LSST, which is a project that sounds like it's ultimately going to happen. If NASA does not pay for it, it will happen more slowly, but it's going to happen.

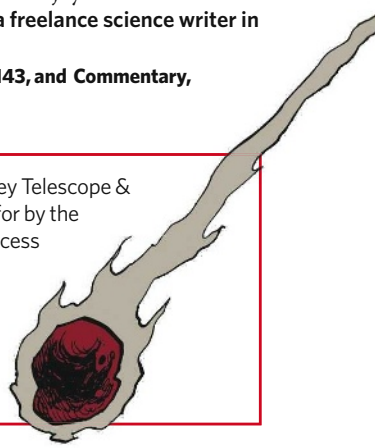
Shoemaker: I liked looking at the sky. I liked doing all those other things that some people thought were awfully time-consuming and might be tedious, but which to me were fun. I retain my interest in all the results, but I don't do the work anymore. For me, the romance of observing is gone.

One of the real pleasures of our programme was the fact that if we found something exciting, and needed confirmation or more positions, we could ask people here and there throughout the world, and they would do follow-up work for us. And by the same token we would do the same thing for them. It was both competitive and cooperative. It doesn't always happen in science, but it worked pretty well for us in those days, and that was a real joy. ■

David Chandler is a freelance science writer in Massachusetts.

See Editorial, page 1143, and Commentary, page 1178.

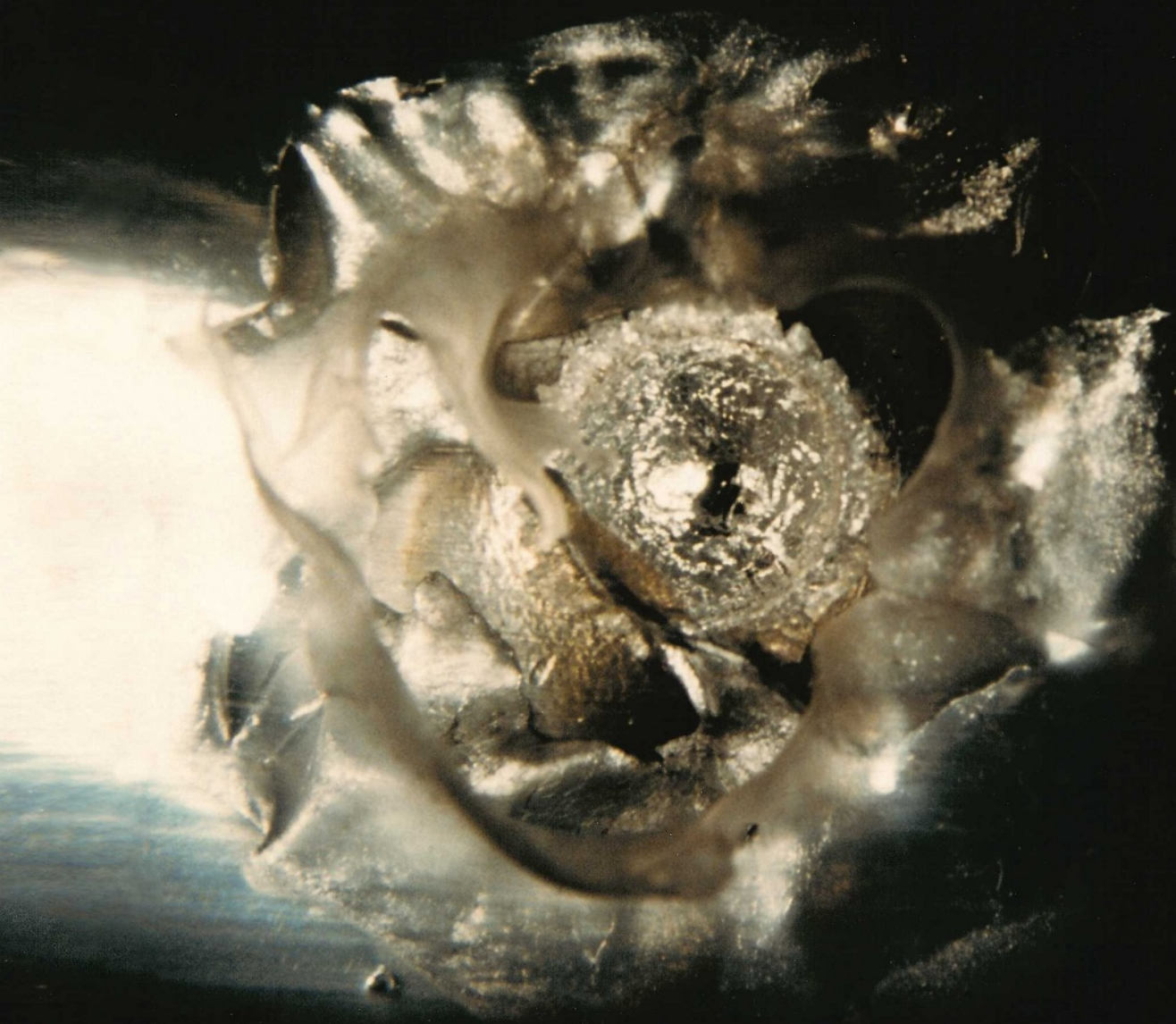
is an expert in orbital mechanics from the University of Pisa, Italy. ● **The B612 Foundation**, named after the asteroid in Saint Exupéry's *Le Petit Prince*, is a think-tank and lobby group set up by Schweickart to further the Earth's protection against asteroid strikes. Its short-term aim is to demonstrate a technological capacity for changing the orbit of a near-Earth asteroid. One technology under discussion is that of a **gravity tractor**, a spacecraft that uses its gravitational attraction to change a small asteroid's trajectory. Surprisingly, such a spacecraft does not need to be very massive, provided that the asteroid and deflection needed are small. ● **Pan-STARRS** is the Panoramic Survey Telescope & Rapid Response System, a telescope array paid for by the Air Force. The first of its four mirrors is in the process of being commissioned. The **LSST**, or Large Synoptic Survey Telescope, is a next-generation instrument that will observe all manner of transient phenomena, including passing asteroids, when it goes into commission in the mid 2010s. **D.C.**



NATURE|Vol 453|26 June 2008

ALL CRATERS GREAT AND SMALL

From a 5-millimetre dent on a satellite to a 3-kilometre pit in the surface of Mars, the scars of impact events can be seen at every scale. We present a gallery of some particularly appealing ones from Earth and beyond.



LDEF

BIG BOY, LONG DURATION EXPOSURE FACILITY

C. KOEBERL

BOSUMTWI, GHANA

NASA

P. L. KRESAN

METEOR CRATER, ARIZONA

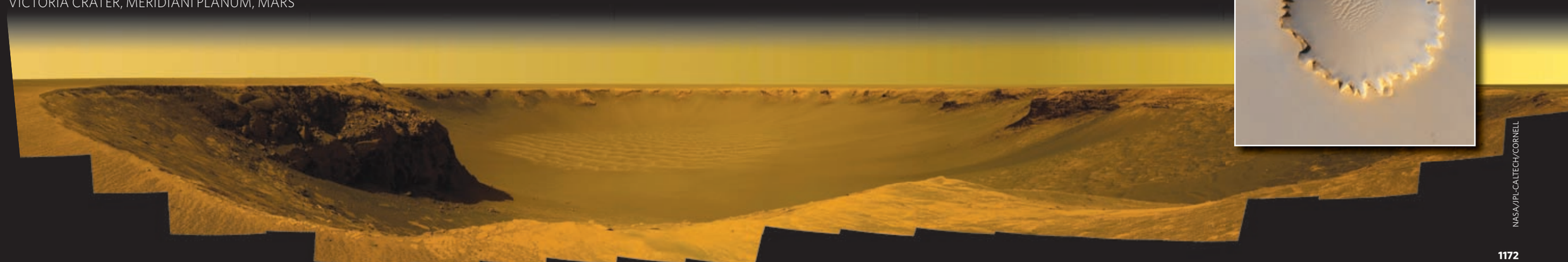
It is unusual for a Chevrolet Malibu to be hit by a meteorite, as Michelle Knapp's was in 1992; but the surface of a satellite such as the Long Duration Exposure Facility, which was in orbit for six years, can expect a pounding. And so can the surface of Earth. Meteor Crater in Arizona, which at 50,000 years old is relatively young and nicely preserved in the desert climate, was one of the first craters to have its impact origin recognized. Many others have since joined it, such as 10-kilometre-wide Bosumtwi, a million years old, which contains Ghana's only natural lake. Human experience of cratered landscapes has now extended to other planets, as shown by this panorama of Victoria on Mars, produced by the rover Opportunity before it ventured into the crater's depths.

PEEKSKILL, NEW YORK

P. THOMAS/LST/ENS LYON

NASA/JPL-CALTECH/UNIV. ARIZONA/
CORNELL/OHIO STATE UNIV.

VICTORIA CRATER, MERIDIANI PLANUM, MARS



NASA/JPL-CALTECH/CORNELL

ZUMBA, DAEDALIA PLANUM, MARS

STICKNEY, PHOBOS

ENKI CATENA, GANYMEDE

NASA/SPL

ZUMBA: NASA/JPL/UNIV. ARIZONA; STICKNEY: NASA/JPL-CALTECH/UNIV. ARIZONA; HERSCHEL: NASA/JPL/SPACE SCI. INST.; CALLISTO: NASA/JPL; EUCLIDES: PHOTO: K. MATTINGLY, APOLLO 16, 16–27 APRIL 1972. FROM M. LIGHT *FULL MOON* (1999)

Large impacts can change planetary surfaces profoundly. The rings that surround the Valhalla basin on Callisto, a moon of Jupiter, take up roughly one-tenth of its surface. The basalts that fill lunar basins, such as those in Oceanus Procellarum, give Earth's satellite its 'seas' — themselves peppered with smaller craters, such as Euclides. A large enough impact can completely destroy its target; those that left the 9-kilometre-wide Stickney crater, on Mars's moon Phobos, and the 130-kilometre-wide Herschel, on Saturn's moon Mimas, came close.

Sometimes, though, it is the impactor that comes to pieces: crater chains such as Enki Catena on the jovian moon Ganymede were probably formed by comet fragments crashing in one after the other. Zumba crater on Mars has a special claim to fame: it is possible that this relatively recent impact threw up into space some of the rocks that have now fallen to Earth as martian meteorites.

VALHALLA, CALLISTO

HERSCHEL, MIMAS

EUCLIDES, OCEANUS PROCELLARUM, THE MOON

CORRESPONDENCE

Agronomy and plant breeding are key to combating food crisis

SIR — In your Editorial 'A research menu' (*Nature* **453,** 1–2; 2008), you highlight the need to spend more on agricultural science to overcome today's food crisis. But this is not just a matter of greater expenditure — the way in which the money is spent is also important.

Reduced public spending on agricultural research might well be partly to blame for the present crisis. But it is also true that what funding there is has increasingly been directed towards molecular aspects of plant growth and development, arguably at the expense of practical agronomy and breeding issues.

Progress in understanding plant molecular biology is impressive, and useful applications are evident when the trait in question is relatively simple. But improvements in yield and input efficiency — essential for sustainability — stem almost exclusively from traditional breeding and agronomy.

Molecular tools such as gene-expression and metabolite profiling are a long way from becoming incorporated into selection procedures for complex-trait breeding. There are many well-characterized quantitative trait loci (stretches of DNA closely linked to the genes that underlie a trait) that affect yield, but there are no clear examples of such loci being successfully backcrossed into high-performance crops.

Diverting most of the limited agricultural-research resources into molecular biology makes it harder to obtain routine funding to improve traditional breeding and agronomy. So, although we agree that a marked increase in funding is necessary, funding bodies should be aware that preferred allocation to molecular biology risks delaying progress on the pressing issue of improving crop productivity.

In this context, European reluctance to fund research into yield improvement has important implications for European as well as developing countries. As the 'green revolution' example shows, such knowledge is useful beyond geographical frontiers.

Lucas Borrás Departamento de Producción Vegetal, Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, S2125ZAA Zavalla, Santa Fe, Argentina
Gustavo A. Slafer ICREA, Catalonian Institution for Research and Advanced Studies and Department of Crop and Forest Sciences, University of Lleida, Centre UdL-IRTA, Av. Rovira Roure 191, 25198 Lleida, Spain

The ethical basis of the null hypothesis

SIR — Further to T. Häusler's 'In Retrospect' review of Sinclair Lewis's 1925 novel *Arrowsmith* (*Nature* **453,** 38; 2008), the book was required reading for graduate courses in professional practice and ethics in the biological sciences that I taught in the 1990s. Arrowsmith's ethical dilemma was whether he should deny some villagers his phage therapy so that they could serve as controls. His conundrum endures to this day — the choice between bequeathing knowledge from a properly designed controlled experiment and risking the health of members of the control group by withholding potentially beneficial treatment.

The control group provides an unbiased test of the null hypothesis, which predicts what to expect if our ideas of how nature works are wrong. It could be argued that it is therefore an ethical obligation for the scientist to take the null hypothesis seriously. No other professional is ethically obliged to consider what might happen if he or she is wrong.

Arrowsmith's employers put the entire ethical burden of choosing the control group and implementing the experiment onto his shoulders. Ethical burdens, however, are properly borne by the entire community. Current practices of having institutional review boards to oversee experiments, obtaining consent from the treated patient, double-blind procedures and full disclosure combine to ensure that the ethical burden of research is shared by all the people involved in the work and is not unfairly placed on a single individual.

John Pastor Department of Biology, University of Minnesota–Duluth, Duluth, Minnesota 55812, USA

Stem-cell urological treatment was not carried out illegally

SIR — I want to express my displeasure at the News story 'Doctors accused of doing illegal stem-cell trials' (*Nature* **453,** 6–7; 2008).

A prospective cohort study and a prospective randomized trial of our procedure for treating urinary incontinence with autologous myoblasts and fibroblasts were filed with the ethics committee of the University of Innsbruck in February 2001. I cannot explain why these documents are no longer intact in the committee's records. Results of the prospective randomized trial, involving 63 women, were published last year (H. Strasser *et al. The Lancet* **369,** 2179–2186; 2007).

The ethics committee itself emphasized that approval for the study should be applied for at the *Arzneimittelbeirat* (pharmaceutical committee) of the Austrian Ministry of Health, owing to the novelty of the projects. After receiving the protocols, the *Arzneimittelbeirat* had no objections to performing clinical studies in June 2002. Andreas Scheil's assertion that the last time approval for studies was given by the government, and not by an ethical committee, was during the Third Reich is insulting.

After our publication in *The Lancet*, the ethics committee asked the director of the Ministry of Health's office for public health, Hubert Hrabcik, to reinvestigate the application and approval for the trial published in *The Lancet*. In October 2007, Hrabcik informed the ethics committee that the *Arzneimittelbeirat* dealt with the trial published in *The Lancet* and confirmed that approval by the *Arzneimittelbeirat* represents ethical approval.

Because of results showing advantages compared with other therapies, and as production of the stem cells required for the treatment of urinary incontinence had been approved by the Ministry of Health, the department of urology decided to offer this therapy outside clinical trials to selected patients who had signed an informed consent. The patient mentioned in your article, Dieter Bollmann, had no postoperative side-effects or complications. I want to emphasize that, although your article states that "patients" are taking legal action, I only know of this single case of a patient suing the hospital.

On 8 May, TILAK (the company that manages the Innsbruck University hospital) and the department of urology of the Medical University of Innsbruck issued a joint statement in which they expressed regret for past misunderstandings, differences of opinion, problems in coordination of procedures, and irritations. A new, extensive clinical study will be performed in close coordination with the ethics committee in reaction to the new European Union directive on advanced therapies.

After the completion of successful clinical studies, it is planned to offer the injection of autologous myoblasts and fibroblasts as standard therapy in the future, to the benefit of the Tyrolean population.

Hannes Strasser Universitätsklinik für Urologie, Medizinische Universität Innsbruck, Anichstrasse 35, 6020 Innsbruck, Austria

COMMENTARY

What Spaceguard did

A survey of large objects near Earth has shown that there is little risk of a cataclysmic impact in the next century. **Alan Harris** asks if such cataloguing efforts should continue.

The sky isn't falling, but there are still good reasons for keeping an eye on it. In 1991, a NASA-sponsored international working group convened to develop a thorough survey of near-Earth objects (NEOs) — predominately asteroids with an orbit that brings them within 1.3 astronomical units of the Sun. The objective of the survey would not be a mere sampling of the large asteroids that might constitute a risk to Earth, but rather a census. The report[1] defined the 'Spaceguard Survey'. Spaceguard's goal was to identify most NEOs larger than 1 kilometre in diameter within a decade. The impact of an asteroid larger than 1 kilometre in diameter has the potential to cause a global climatic perturbation, similar to a 'nuclear winter', and could lead to billions of deaths worldwide[2]. Such events, although less frequent than smaller 'impacts' such as the Tunguska event in Russia (see page 1157), nevertheless present a greater risk of death, even to individuals. Moreover, they carry the additional risk of ending civilization. So, it is clear to most why a survey might be important.

The idea was slow to catch on within NASA, but by May 1998, Carl Pilcher, the Science Director of Solar System Exploration in the NASA Office of Space Science, testified before the Subcommittee on Space and Aeronautics of US Congress that "NASA is committed to achieving the goal of detecting and cataloguing 90% of NEOs larger than 1 kilometre in diameter within ten years". This for many was the formal start of Spaceguard, so it is appropriate, a decade later, to ask whether its goals have been met. The pedantic answer is no, but in terms of risk reduction — or more precisely, knowing whether an impact will, or will not, occur in our lifetimes — Spaceguard identified a fraction of NEOs responsible for more than 90% of the potential impact risk, and found that impacts from that fraction pose a negligible risk in the next 50–100 years. The remaining short-term risk is almost entirely from any remaining undiscovered NEOs. In that sense, the Spaceguard Survey has been a remarkable success.

Two years ago, I was commissioned by NASA, through the NEO Program Office at the Jet Propulsion Laboratory in Pasadena, California, to assess the progress of Spaceguard. I filed my final report with NASA in March 2007 and have presented a brief summary of the main results[3]. It is easy enough to keep count of the number of discovered objects larger than a given size, but to know when 90% have been found, one must estimate the total population. This is a bit of a bootstrap process, using the survey itself to estimate everything out there. In the simplest terms, if we scan the sky tonight, the number of detections of already-known objects compared to the total number of objects detected during a test interval gives us a measure of completeness[4]. In detail, it is not so simple because not all NEOs are equally detectable.

How is Spaceguard doing? As of 10 June 2008, 742 near-Earth asteroids of diameter greater than 1 kilometre had been discovered[5]. In my report, summarized in Figure 1, I estimated a total of 940, and so the Spaceguard Survey has identified about 79%; not quite 90%, but not bad considering the uncertainties and the efforts required to reach 90%.

Meanwhile the estimated risk of impact is dwindling. In the very largest size range, asteroids about 10 kilometres in diameter, the three already discovered are almost certainly all that exist. These would produce an impact similar to that which killed the dinosaurs 65 million years ago, with an estimated impact interval of around 10^8 years — roughly the last time dinosaurs walked on Earth. Oddly, an object that might cause a Tunguska-like event — roughly 50 metres in diameter — should collide with Earth only about every 1,500 years, and the last event we saw was only 100 years ago.

Recently, Mark Boslough at Sandia National Laboratories, in Albuquerque, New Mexico, suggested that the energy of the Tunguska event may have been as low as 3 megatonnes[6]. That adjustment reduces the expected time between similar events to perhaps about once in 500 years, still leaving the chances of an event within a century as unlikely. 'Statistics of one' cannot be held too rigorously to formal probability estimates, but our view of the skies has produced a strong predictor for the frequency of impacts. It is so strong, in fact, that it could and should rule out some suggestions of past impacts such as the multiple kilometre-sized objects claimed by some to have pelted Earth during the Holocene period[7]. Such an event is inconsistent with what we see in the skies, by about two orders of magnitude.

Another NASA study[8] in 2003, estimated the expected damage from impacts of various sizes. Using those values of expected damage, and the impact frequency from the newly derived population (Fig. 1), I estimated the 'risk spectrum' of impacts over the entire size range of those that can penetrate the atmosphere. Figure 2 shows that 'spectrum', first for the entire population, that is, the 'intrinsic risk' before any NEOs had been discovered, and secondly the 'residual risk' from the fraction of the NEO population that remains undiscovered. Since the objects that have been discovered have been found to have no, or a vanishingly small, probability of hitting Earth in the next 50 or more years, we can think of that fraction of the intrinsic risk as 'retired' for the short term over which we can predict impact trajectories, about a human lifetime.

Figure 2 shows that the risk from large impacts — the

kind that would cause global climatic disaster and potentially bring down our civilization — has been dramatically reduced, by more than an order of magnitude. In the smaller size range, from several-hundred-metre-diameter objects that could cause massive tsunamis if they crashed into an ocean, down to sub-hundred-metre objects the size of that in the Tunguska event — which could cause ground damage from airbursts — current surveys have done little to retire the risk. But the intrinsic risk from these events is very small, and in fact resembles that of other natural disasters such as tsunamis, earthquakes and volcanic eruptions in that they do not pose a global threat to life as we know it.

In the 2003 NASA report[8], the recommendation was made for a new survey to reduce the assessed residual impact risk from objects less than 1 kilometre in diameter by a further order of magnitude. It was estimated at that time that to achieve this goal would require discovering 90% of NEOs larger than 140 metres in diameter. This has become the new mantra of survey plans[9], but perhaps this should be reconsidered. Because of the steep dip in the population curve in the size range between about 50 metres and about 500 metres, the intrinsic impact frequency, and hence the impact risk, is about three times lower than was estimated in the 2003 report. So, in a way, two-thirds of the risk assumed to exist in those reports is gone already, without even looking at the sky. In the earlier reports, the 'residual risk' to be addressed by a next-generation survey was assumed to be approximately 300 fatalities per year, but using my new population estimate that figure drops to around 80 per year. In comparison to other risks in life, this is negligible.

What is the risk that your death will come from the sky? Before the Spaceguard Survey, it was thought to be comparable to the risk of dying in a commercial aeroplane accident. Currently, however, the residual risk from the remaining undiscovered NEOs is more comparable to the risk of death from a fireworks accident (see graphic, previous page). At some point one has to ask how far down we need to drive the residual risk, especially because the cost of doing so increases steeply as the size of impactors decreases.

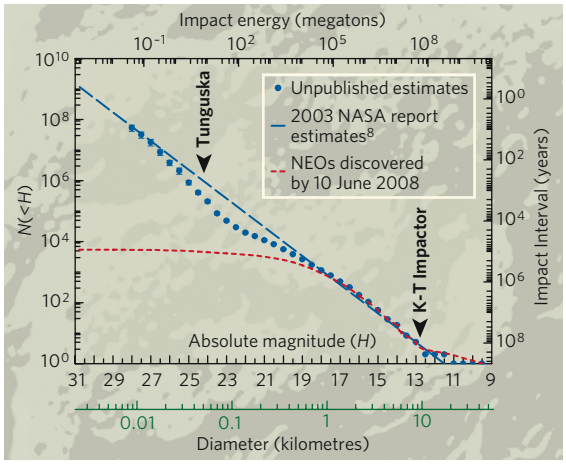


Figure 1 | Estimate of the cumulative population of near-Earth objects (NEOs) versus size. *H* is the absolute magnitude (brightness at standard distance of 1 astronomical unit from Earth and Sun), and $N(<H)$ is the cumulative number of objects with *H* less than a given value. The fraction currently detected is nearly complete to $H \sim 16$, but falls off rapidly with increasing *H* magnitude. Other scales are derived. For example: diameter is inferred from mean reflectivity, impact energy and impact interval require additional knowledge of orbital characteristics.

"In comparison to other risks in life, the risk of death from a small impact is negligible."

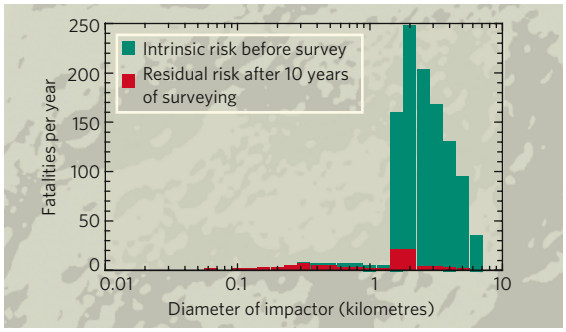


Figure 2 | How the Spaceguard Survey has reduced the short-term risk of impacts from near-Earth objects.

Be that as it may, plans are continuing for next-generation surveys and they may serve another purpose. The Large Synoptic Survey Telescope (LSST), a ground-based, wide-field instrument with an 8.4 metre aperture, is planned to enter service by about 2012. Recognizing the diminishing value of driving our assessement of the impact risk so low, the LSST project has adopted the NEO survey as only one of many scientific goals for the telescope, and in particular has emphasized the scientific value of a Solar System survey.

Indeed, the Spaceguard Survey itself has yielded scientific results aside from the value of impact risk reduction. Recently William Bottke and his colleagues at the Southwest Research Institute in Boulder, Colorado, used orbital statistics of asteroids discovered by the surveys to propose that the event that killed off the dinosaurs came from an 'impact shower' resulting from the collisional breakup that produced the Baptistina asteroid family[10]. The size-frequency distribution of impactors (Fig. 1) is itself interesting. The drop in numbers from those of a few hundred metres in diameter to those of a few tens of metres is not yet explained, but is perhaps due to the transition, at around 200 metres diameter, from 'rubble pile' structure among larger asteroids, which are less resistant to disruption by collisions, to monolithic bodies in the smaller size range, which are more resistant to further collisional breakup[11]. Thus, although continuing surveys for the sole purpose of risk reduction may be of diminishing value, the scientific rewards will remain high, and we can hope that ever more powerful surveys will continue in the future. ■

Alan Harris is a senior research scientist with the Space Science Institute, 4603 Orange Knoll Avenue, La Canada, California 91011-3364, USA.

1. Morrison, D. *The Spaceguard Survey: Report of the NASA International Near-Earth-Object Detection Workshop* (NASA, 1992).
2. Toon, O. B., Zahnle, K., Morrison, D., Turco, R. P. & Covey, C. *Rev. Geophys.* **35,** 41–78 (1997).
3. Harris, A. W. *B. Am. Astro. Soc.* **39,** 511 (2007).
4. D'Abramo, G. *et al. Icarus* **153,** 214–217 (2001).
5. http://neo.jpl.nasa.gov/stats/
6. Boslough, M. Am. Geophys. Union, Fall Meeting 2007, abstract U21E-03.
7. Pinter, N. & Ishman, S. E. *GSA Today* **18,** 37–38 (2008).
8. http://neo.jpl.nasa.gov/neo/neoreport030825.pdf
9. http://www.b612foundation.org/papers/NASA-finalrpt.pdf
10. Bottke, W. F., Vokrouhlický, D. & Nesvorný, D. *Nature* **449,** 48–53 (2007).
11. Pravec, P., Harris, A. W. & Warner, B. D. in *Proc. Int. Astron. Union Symp.* (eds Milani, A. *et al.*) **2,** 167–176 (Cambridge Univ. Press, 2006).

See Editorial, page 1143, and News Features, pages 1157 and 1165.

BOOKS & ARTS

The end of the line?

A spotlight on the historic US fishing port of Gloucester fails to capture the complexity of the fisheries collapse caused by overexploitation and regulation, says **Daniel Pauly**.

The Last Fish Tale: The Fate of the Atlantic and Survival in Gloucester, America's Oldest Fishing Port and Most Original Town
by Mark Kurlansky
Random House/Jonathan Cape: 2008.
304 pp. $25.00/£16.99

NEWSCOM

Net profit: efficient trawler technology has led to dwindling stocks of fish such as cod.

In his new book, Mark Kurlansky follows a formula that has served him well in earlier works on cod, salt, the Basques and oysters: pick a seemingly mundane maritime topic, dig deep into the historical archive for savoury anecdotes, add a sprinkling of cooking recipes and serve it up with a *bon-vivant*'s style.

The Last Fish Tale is the story of Gloucester, Massachusetts, the oldest fishing port in the United States. Kurlansky spotlights this New England town to investigate the decline of Atlantic fisheries. He describes Gloucester's fascinating history, a product of its insularity and island geography, its strong egalitarian identity and the large number of fishermen, drawn from a succession of immigrant communities, lost at sea. With rich ingredients and engaging writing, the book should work. Readers might agree that the loss of yet another diverse, insular culture is bad. But Kurlansky listened to too few voices, and his resulting picture is unbalanced.

My confidence was shaken early in the book. Kurlansky tells us that, in 1602, the explorer and privateer Bartholomew Gosnold remarked that "the fish were far bigger [in New England] than those in the north". The author repeats this fact throughout the book, even though Gosnold is apparently its only source. Twentieth-century ichthyologists demonstrated that the opposite is true. By studying the maximum sizes of various fishes, they showed that fish grow larger, all other things being equal, in the colder waters at the poleward ends of their range. This error matters: sources must be checked against others to avoid drawing the wrong conclusions.

After describing the town and its denizens, the author explains how Gloucester ran out of fish, especially Atlantic cod. The decline of this once-abundant species was partly caused by the success of the schooner-based fishery, which, even though it relied on wind power, harvested enough to reduce the stock. Bottom trawlers dealt the *coup de grâce*. Kurlansky recalls the introduction of the murderous trawling gear in Gloucester where, as elsewhere, it was first viewed with suspicion, then adopted because its effectiveness was irresistible. This simple explanation should suffice: the cod declined because of overfishing.

Yet Kurlansky demurs, and hints darkly at other causes. When we accompany him to Newlyn, a fishing town in Cornwall, UK, which he presents as Gloucester's Old World doppelgänger, we meet fisheries regulators who cannot tell a bass from a cod. "Newlyn vessels had been landing more than their quota of cod, hake, and monkfish by labelling them ling, turbot, and bass — fish for which there were no quotas," he states. That it took five years for the regulators to discover this, Kurlansky says, indicates how little they know about fish. Yet it is just as likely that these officials were tolerating an illegal practice, as is common in fisheries worldwide.

Like the Gloucester fishermen, Kurlansky believes that bureaucrats from the US National Marine Fisheries Service cause the problems, not fishing practices. The stocks may have disappeared but the fishermen have not, and everybody is looking for the crumbs of a vanished pie. Although the author tells us at length about the antics of the fishermen at Gloucester harbour festivals, such as competitive scrambles along a greasy pole, he does not tell us how, in that same harbour, two fisheries regulators were hanged in effigy in 1999. These officials wanted only to reduce the pressure on vanishing stocks, prevent further declining resources, and keep the fisheries going.

As Kurlansky's informants did not deliberately mislead him, this case does not mirror that of anthropologist Margaret Mead misreporting on the sexual mores of Samoan youths. Rather, it is a case of shared delusion, similar to that of John Edward Mack, the Harvard University psychiatrist who studied people who believed they had been abducted by aliens. Adopting his subjects' obsessions, he wrote a book arguing that cosmic kidnap was real.

These are strong words, particularly as I liked and learnt from Kurlansky's previous books. But *The Last Fish Tale* fails to explain the dual roles of the fishermen as both victims and ferocious drivers of the overfishing

behind the collapse of the Gloucester and New England fisheries. Until we reveal these dual roles and the ensuing pathologies, there will be no rebuilding, no renewal of the fisheries.

I suspect that this book, ironically, will find popularity among the tourists who flock to a gentrified Gloucester. Under Kurlansky's disapproving gaze, they will gradually displace the fishermen, as in most fishing towns around the north Atlantic. Visitors to Gloucester will love the book and the town's many charming features described in its pages. They will think of the fish and shake their heads at such a loss, still failing to understand. ■

Daniel Pauly is professor of fisheries and director of the Fisheries Centre at the University of British Columbia, 2202 Main Mall, Vancouver, British Columbia V6T 1Z4, Canada.

Making genetic history

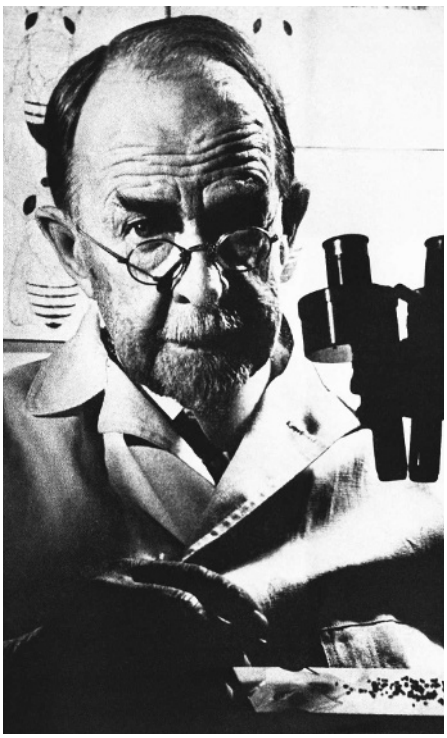
In Pursuit of the Gene: From Darwin to DNA
by James Schwartz
Harvard University Press: 2008. 384 pp.
$29.95, £19.95, €22.50

When I was a student, 'doing genetics' meant crossing two different strains or species. Now it means sequencing DNA, preferably human. Between these two poles lies the history of genetics, a pathway fraught with sharp turns, steep gradients and dead ends — and engagingly recounted in James Schwartz's new book.

Despite its subtitle, *In Pursuit of the Gene* is not a comprehensive history of genetics, but focuses solely on classical genetics. Schwartz, a science writer, begins with Charles Darwin's ill-fated 'pangenesis' theory of the inheritance of acquired characteristics, and runs through the rediscovery of Gregor Mendel's work on inherited traits. The story continues with the consolidation of Mendelism and chromosomal inheritance by Thomas Hunt Morgan and his students in the 'Fly Room' lab at New York's Columbia University, where modern genetics began, and concludes in 1946 with Hermann Joseph Muller's Nobel Prize in Medicine for inducing mutations with X-rays. Later history, from the discovery by Oswald Avery and colleagues that DNA was the 'transforming principle', to the Human Genome Project, is squeezed into a 12-page epilogue. Those seeking a history of molecular genetics should read Horace Freeland Judson's magisterial *The Eighth Day of Creation* (Simon & Schuster, 1979).

Many histories of genetics cover the same ground. What distinguishes Schwartz's account is his impeccable scholarship, based on many primary sources, and his ability to keep the narrative moving, interweaving discoveries with the strong and eccentric personalities who made them. He does not slight the science, describing experiments in detail so dense that the reader is advised to keep a pencil and paper handy. The effort required to understand the book may, sadly, remove it from the ambit of popular science.

The book's apogee is its tale of the "Mendel Wars" around the beginning of the twentieth century, the struggle to bring together Mendel's ideas on heredity and Darwin's theory of evolution. On one side were the Mendelians, including Francis Galton, William Bateson and Charles Hurst, who accepted Mendelism but considered natural selection as ineffective, seeing evolution as occurring by 'macromutations', or single genetic changes of very large effect. On the other side stood the biometricians, most notably Karl Pearson and Raphael Weldon, who accepted the ubiquity of Darwinian selection but rejected Mendelian genetics. Given the strong egos involved and the fundamental nature of the science at stake, the battles Schwartz recounts were fierce. Friendships were destroyed, careers threatened. After a particularly contentious meeting about the genetics of horse coat colour at the Royal Society in London, Pearson hissed at Hurst, "You shall never be Fellow here as long as I live".



BETTMANN/CORBIS

Fruitful collaborations were formed in Thomas Hunt Morgan's fly genetics lab.

Other high spots in the book include the early and now largely forgotten work on cytological genetics by Walter Sutton and Edmund B. Wilson, involving years of eye-strain from squinting at confusing chromosomal preparations of sea urchins, aphids and grasshoppers. These studies established that different chromosomes carry different hereditary factors, yet occur in pairs that become separated during the formation of gametes in meiosis, giving essential physical support for Mendel's laws.

The book's longest section details the immense contributions of research on the fruitfly *Drosophila melanogaster* to our understanding of heredity. Schwartz explains how, from 1912 to around 1930, Morgan and his 'boys', Alfred Sturtevant and Calvin Bridges, along with Muller, were "responsible for the integration of Mendelism and the chromosome theory that is the basis of genetics". Within a few years, this conjunction of remarkable intellects in a tiny laboratory led to methods for mapping chromosomes both genetically and cytologically, and to the discovery of sex linkage, chromosome inversions, nondisjunction and many other phenomena that now form the dogma of transmission genetics.

Alas, here we find a major flaw. Schwartz notes that he was inspired to write his history by reading Elof Carlson's worshipful biography of Muller, *Genes, Radiation, and Society* (Cornell University Press, 1981). But this only generates further hagiography: the discussion of Muller's work occupies a quarter of *In Pursuit of the Gene*, a disproportionate chunk. Schwartz gives the impression that Muller, or ideas purloined from him by others, was behind nearly every advance in fly genetics. Sturtevant's contributions are given short shrift, Morgan is portrayed as a conniver who acquired his Nobel status on the backs of his students, and Bridges — perhaps the finest pair of eyes ever to peer at a magnified fly — is dismissed as being "famous for stealing other men's wives as well as their ideas". Schwartz does not mention the work of Lewis Stadler, who independently discovered X-ray induction of mutations in barley at the same time as Muller's work on *Drosophila*. Like many plant geneticists, Stadler was marginalized as a glorified crop breeder.

It is easy to sympathize with Muller, who had a tumultuous life and was the perennial underdog: Jewish, short, bald and with a high voice.

Fractious, and possessed of unpopular socialist views, he floated from university to university, winding up in the Soviet Union until he fled to escape Trofim Lysenko's destruction of Russian genetics. Yet during all these peregrinations he maintained an uninterrupted programme of research. It is a scandal that Muller did not secure a tenured academic job until he was 55 — he won the Nobel prize a year later.

Muller was one of the best geneticists of the twentieth century, a visionary who predicted the rise of molecular genetics and the use of association mapping to identify genes for human behaviours. He was also difficult to work with, obsessed with credit and depressive to the point of once attempting suicide. Schwartz repeatedly states that Sturtevant, Bridges and Morgan tried to ruin Muller's reputation by stealing his ideas and slandering him, but the evidence is unconvincing. Working together in the Fly Room, talking science as they worked on flies in what was a continuous lab meeting, it is not surprising that they shared ideas and information. After all, it was Sturtevant who gave Muller the idea of using lethal alleles to measure mutation rates.

The other 'boys' were not slouches. Bridges discovered nondisjunction, thereby proving the chromosomal theory of heredity, and published it as the first paper in the first issue of the journal *Genetics*. He constructed the first map of genes on autosomes, did fundamental work on sex determination and produced maps of *Drosophila* salivary-gland chromosomes that have never been bettered. Sturtevant was the first to establish, while still an undergraduate, that genes are arrayed linearly on chromosomes. He devised the chromosomal fate mapping later used so effectively by the geneticist Seymour Benzer, founded *Drosophila* taxonomy and, by studying the action of eye-colour mutations in the fly, became the father of biochemical genetics. But neither Sturtevant nor Bridges was obsessed with priority: Sturtevant was the most modest of men, whereas Bridges, a great womanizer, had more pressing interests.

In Pursuit of the Gene should be required reading for all biologists unfamiliar with the history of genetics. Schwartz shows how quickly science can advance when a group of first-class minds encounters a fertile but unploughed field. Progress in genetics, as in all modern science, was truly a collaborative affair. There was no Darwin of genetics — not even Muller. There was, and is, plenty of credit to go around. ■

Jerry A. Coyne is a professor in the Department of Ecology and Evolution at the University of Chicago, Illinois 60637, USA.

Swayonomics

Sway: The Irresistible Pull of Irrational Behavior
by Ori Brafman and Rom Brafman
Currency/Doubleday: 2008. 224 pp. $21.95

In the Biblical parable in Matthew 25:14–29, a servant who was given five talents of money invested them and returned ten talents, whereas a servant given one talent buried it in the ground without profit. The master gave the risk-averse servant's one talent to his successful rival. The effect was elevated into a principle: "For to everyone who has, more shall be given, and he will have an abundance; but from the one who does not have, even what he does have shall be taken away."

Sometimes named the 'Matthew Effect', marketers call this response 'cumulative advantage'. I think of it as the 'bestseller effect'. Every author and publisher knows that once a book gets a head-start in sales it signals to consumers that other people want that book, causing them to desire it and purchase more, so the richest authors get even richer.

In *Sway*, the brother authors Ori Brafman, an entrepreneur, and Rom Brafman, a psychologist, describe the social and psychological effects that shape our beliefs and behaviours. They hope to trigger their own Matthew Effect with this highly readable book. But predicting the next bestseller is as reliable a business as astrology. That problem affects all books, including, ironically, those about marketing and behaviour: the psychological principles may explain what happened in hindsight, but cannot be used to predict the future.

Sway is a fun read, and the brothers Brafman are compelling storytellers, pulling in the reader immediately and narrating at a breezy pace. But the book is thin on science and thick on anecdotes. The authors have a propensity for 'just-so' stories, favouring this or that behavioural principle when other explanations exist.

The book opens, for example, with the tragic 1977 crash of KLM flight 4805 during take-off from the tiny Tenerife airport in the Canary Islands. While motoring down the runway, the Boeing 747 slammed into Pan Am flight 1736, also a 747. The crash was the worst disaster in aviation history. What was the cause? The authors argue that it was psychological. The KLM captain Jacob Veldhuyzen van Zanten was a top pilot, featured in airline advertisements, who took pride in getting his passengers to their destination on time. That day he was behind schedule, having been rerouted to Tenerife after a bomb threat at his destination airport, and delayed on the island by fog. Captain van Zanten worried about his reputation for punctuality. "An unseen psychological force was at work," claim the authors, "steering van Zanten off the path of reason." This force was "loss aversion". Behavioural economists have shown that when we make a decision, potential losses hurt twice as much as potential gains feel good. "This principle is key to understanding van Zanten's actions," the Brafmans explain. He dreaded "the cost of putting up the passengers, the chain reaction of delayed flights and the blot on his reputation for being on time".

"People find evidence for what they already believe and ignore anything contrary."

Baloney. Van Zanten's plane was one of several large aircraft diverted to Tenerife. They manoeuvred tightly around the runway, the taxiway that ran parallel to it and four small connector taxiways between the two. Several spilled over onto the taxiway, so some planes had to taxi up the runway, turn around, and then take off down that same runway. Van Zanten did this, but after turning around in preparation for take-off, the fog reduced visibility to 300 metres. Unknown and invisible to van Zanten, at the same time Pan Am 1736 had been instructed to taxi down the same runway and take the third exit on its left in order to avoid the KLM flight's take-off. After clarifying which exit to take — "The third one, sir; one, two, three, third, third one" the controller emphasized — the Pan Am jet counted them off against an airport diagram. The cockpit voice recorder revealed that the Pan Am crew identified the first two connecting taxiways, but missed the third; the collision happened near the fourth exit.

Meanwhile, in the KLM plane, van Zanten's co-pilot radioed the tower for clearance. The tower did not clear them for take-off immediately. At this moment, a call from the Pan Am jet to the tower caused interference on the radio. The Pan Am crew signalled that they were still on the runway, but because of the radio interference the KLM crew did not hear the message, and began their fateful take-off sequence. The

airport lacked ground radar so no one could locate the planes. By the time van Zanten saw the Pan Am plane it was too late. He throttled his engines full and pulled up the nose of the plane, but his fuselage clipped the top of the Pan Am jet, ripping it to shreds. The Pan Am pilot hit his engines and turned sharply into the exit path, but it was too little too late. Total death toll: 583.

The cause of this crash, investigators concluded, was a concatenation of conditions, none of which had anything to do with the psychology of loss aversion: bad weather, crowded conditions, big planes on a small runway, and misinterpretations and false assumptions.

Even if we grant the brothers Brafman the option of looking for an 'ultimate' instead of 'proximate' cause of the crash in the form of cognitive biases and behavioural persuaders that drove van Zanten to make his fateful decision to take off, loss aversion would be low on a causal vector list. Top of my list would be the 'confirmation bias', in which people look for and find confirmatory evidence for what they already believe and ignore evidence to the contrary. Once van Zanten thought he got the "OK" for take-off, everything else made sense. Or, perhaps it was the effect of 'inattention blindness', in which people attend to one task so intently that they miss obvious things in their visual field. Or it could be the 'self-serving bias' and the 'better-than-average bias' that made van Zanten overconfident in his abilities and thus less risk-averse than he might normally be. Maybe there was a 'priming effect', such that van Zanten's brain was primed to hear "take-off" in that garbled radio message. Or how about just the power of expectation?

The real problem here is the hindsight bias. Not for van Zanten, but for observers trying to read into a past event psychological effects that have been measured in the laboratory. The research on cognitive biases and judgemental heuristics — cleverly used in the service of reconstructing past events by the authors of *Sway* — is well grounded in empirical data, but the Brafman brothers face the same problem as the rest of humanity in trying to make sense of seemingly chaotic human behaviour: those very same biases operate in the process of using them to explain someone else's behaviour. Call it the 'meta-heuristic' bias. ■

Michael Shermer is the publisher of *Skeptic* magazine, a columnist for *Scientific American* and professor in the School of Economics and Politics at Claremont Graduate University, California. His latest book is *The Mind of the Market*.

A. ENGLEN

Q&A: Travels with a paintbrush

Watercolour artist and explorer **Tony Foster** paints in some extreme places. He has climbed mountains, sketched erupting volcanoes and drawn underwater. As an exhibition of his works of Mount Everest and the Grand Canyon opens in London, he tells *Nature* why he goes to such extraordinary efforts.

Why did you decide to paint remote and dangerous landscapes?

I was a pop artist originally. But I got fed up with using second-hand imagery and thought I should work on things I experienced myself. My first trip followed the journeys of US writer and philosopher Henry Thoreau through the wildernesses in Maine. It seems fairly mundane now. My trips have become more and more extreme.

Your recent paintings are large, yet you paint *in situ*. Does this present unusual challenges?

All the difficulties are magnified by the scale and the location. It's much more laborious to do a big painting than a small one, and difficult physically to haul a 2-metre-wide drawing board around and lash it to the rocks in high winds. At subzero temperatures, the water for my paint freezes so I mix it with gin.

I suffered from altitude sickness in the Himalayas. I didn't realize how ill I was. I got sicker and sicker until I realized I couldn't carry on. I was coughing blood.

Sometimes it is appallingly difficult and miserable. That's spiced by moments of extraordinary joy if things work out.

Natural subjects were traditionally drawn by artists; now photography has taken over. What are your paintings trying to capture?

I'm not striving for accuracy, but honesty. The work looks different if done *in situ*, rather than from a photograph, which doesn't contain enough information. My paintings evoke a much greater emotional response. The work isn't just about how the landscape looks, it's about what it's like to live in it and to take the journey.

My exhibition pictures are framed with maps, diary notes and souvenirs. Flint arrowheads on the Grand Canyon paintings symbolize that it has been inhabited for thousands of years. The souvenirs under the Tibetan painting are Buddhist objects. One is wrapped up in Chinese newspaper, bound up and sealed to symbolize the suppression of Tibetan Buddhism.

How did you approach your painting of the Grand Canyon?

It's like doing an enormous jigsaw puzzle. If you try to push in bits that are the wrong shape, it will never work. Two of my most stalwart hiking companions are scientists, geologist Bill Brace from the Massachusetts Institute of Technology and Winslow Briggs, a Stanford University plant biologist. Travelling through the Grand Canyon with a world-class geologist really made me look.

I don't think art has to have a purpose, but if my work has one then it is to bring back to people these magnificent places of untouched nature that are sublimely beautiful and worthy of our attention and protection. ■

Interview by **Daniel Cressey**, a reporter for *Nature* based in London.

Searching for a Bigger Subject: Tony Foster
Royal Watercolour Society, Bankside Gallery, London
2–20 July 2008; then until September 2009 in various galleries in the United States.

In Retrospect: Lucifer's Hammer

Oliver Morton recalls how the first major science fiction novel to depict an impact event conjured the thrill and the horror of natural cataclysm — and even inspired some researchers.

Lucifer's Hammer
by Larry Niven and Jerry Pournelle
HarperCollins: 1977. 494 pp.

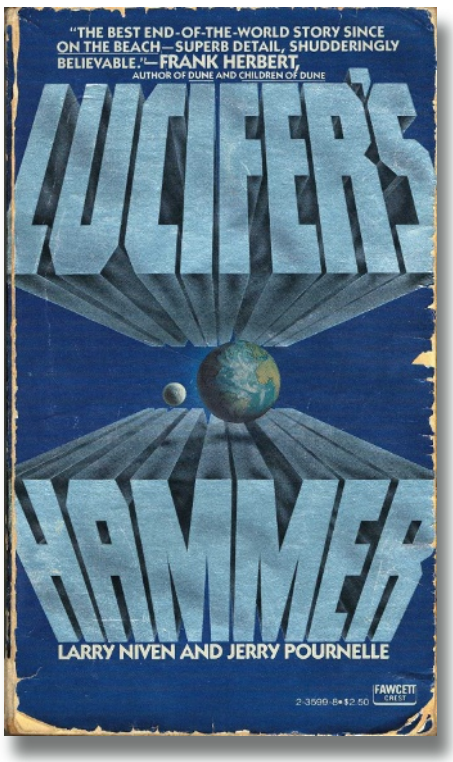
Everyone remembers the surfer. When fragments of comet Hamner–Brown strike Earth a third of the way into *Lucifer's Hammer*, the surfer is floating on his board off Santa Monica, California. The flash in the sky and the fiery cloud on the horizon warn him what's coming. He paddles out to face his death — a tsunami that lifts him to the sky and turns him to the land. He rides the end of the world into the Los Angeles basin like a Hot Tuna Valkyrie, locomotive-fast and skyscraper-high, imagining for a moment that, despite the unbearable strain in his legs and the weight of an ocean at his back, he might still survive to tell his story: "a surfing movie with ten million in special effects!"

The surfer is remembered because he provides the novel's most enduring taste of the sublime — a simultaneous evocation of terror and wonder, immediate danger and cognitive distance, common to disasters imagined and real. As *The Times* columnist Matthew Parris put it, writing with disturbing honesty about the mixed horror and thrill he felt contemplating the Indian Ocean tsunami of 2004: "A minor but insuppressible part of me has almost relished — yes, relished — those huge numbers. As the newspaper headlines spoke greedily of the numbers of dead "approaching" twenty, then fifty, then eighty, then a hundred thousand, something undeniable twitched in the back of my brain ... as though some great auctioneer of calamity were taking bids from the media floor, and I was willing the bidding to carry on upwards .... When the gods themselves strike — then I believe a new depth to our fascination opens."

Published in 1977, *Lucifer's Hammer* was the first major science fiction novel to try to deal realistically with the planetary emergency of an impact event. It plumbs those depths of fascination on an epic scale, rewarded at the time with sales far beyond the normal expectations of the genre.

Authors Larry Niven and Jerry Pournelle originally outlined a story in which aliens, planning to invade, drop a small asteroid on the earth as a useful softening-up exercise. Their editor told them to forget the aliens and concentrate on the asteroid. They did so, making it the centre-piece of a disaster novel carefully fashioned for best-sellerdom, with a large cast and a heft that promises to take up a lot of beach time. In the process they converted the asteroid to a less predictable but more spectacular comet, with doubts about its trajectory allowing a tension-building introduction to their disaster-movie-ready cast of Angelenos.

The novel works well, still, as an airport read. (It is better than the alien-invasion novel the authors finally wrote almost a decade later, although to be fair, *Footfall* (Del Rey, 1985) does have a magnificent space battle at its finale.) And *Lucifer's Hammer* is distinguished by its thorough and informed imagining of a broadly plausible cataclysm never before described. Niven and Pournelle went out of their way to make the impact and its attendant horrors — tsunamis, earthquakes, climate change and crop failure, wars for the best remaining farmland — believable at a time when remarkably little scientific attention had been paid to such things. When the book emerged, it encouraged more of that attention. "*Lucifer's Hammer* killed the dinosaurs," said US physicist Luis Alvarez, saluting the authors when, a few years later, they attended his lecture on the geochemical evidence he and his son had found of a massive impact at the end of the Cretaceous period.



For all that, the book isn't about the comet. It is about the enervating fragility of civilization. Pre-comet, Los Angeles is already perched on the cusp of disaster, as always; in the city's Cielo Drive, Charles Manson has already proved civilization "neither eternal nor safe". "Nothing silly about being ready for the end of civilization," opines a wise biker camping off Mulholland. There's a period pall of sweaty paranoia and doom that gives the impact an almost cleansing feel. At times, the thrill of destruction and survival seems to totter into glee at the settling of scores. Feminism does not outlast the cataclysm, and it is not much missed. Woolly-headed pinko environmentalists eat their enemies and each other, as do most of the book's black characters — a development with disturbing echoes, to say the least, of George Fitzhugh's 1857 antebellum tract *Cannibals All! or, Slaves Without Masters*.

The authors' main theme is that, comet or no, a civilization has the morality its machinery allows it to afford, and that saving the last nuclear power plant is worth a war if it avoids a return to serfdom and slavery. Pangs of survivor guilt and a nostalgia for lost niceties cut across the suspicion that some of the protagonists like the Hammer-scourged world a bit too much. The sacrifice of the scientist who devotes his last winter to making poison gases for the power plant's protection, rather than insulin that would save his own life, is manipulative. But it is plausibly and effectively so. And the richer crops that grow where the gas pooled and corpses fell are a powerful symbol of good from ill. It is hard not to feel a sense of uplift as, on the last page, emblems of industry and trade — electricity, an IOU and, most beautifully, an aircraft's contrail — herald returning life and dignity. ■

Oliver Morton is *Nature*'s Chief News and Features Editor. He is author of *Mapping Mars* and *Eating the Sun*.

See Editorial, page 1143.

D. PLUESS/EPA/CORBIS

Message from the heavens

Discerning the meaning behind Maurizio Cattelan's violent, provocative and now highly valuable sculpture of Pope John Paul II felled by a meteorite raises many questions for viewers, explains **Martin Kemp**.

Pope John Paul II, dressed in his ceremonial regalia, lies prostrate on a rich red carpet. Clinging to his crucifix crozier, he frowns with disquieting intensity, his eyes tightly shut. Nearby lies a scattering of glass shards. A chunky meteorite has plummeted from the heavens, smashed through the gallery skylight, and come to rest in the crook of his bent leg. We presume that the life-size representation shows the pontiff as dead or injured.

What are we to make of this provocative work by the Italian sculptor Maurizio Cattelan? The sculpture is deemed culturally important. It is of high financial value, and was sold to a private collector in 2004 for US$2.7 million. Exhibited in prestigious galleries throughout the world, it uniformly attracts media attention and religious controversy.

We can read the narrative readily, but its meaning is harder to discern. There is a great tradition of death narratives in Italian art. We may see a parallel with the martyrdom of St Stephen, who was stoned to death. Cattelan has been careful with the iconographical details. He has replicated the crozier that was originally made by Lello Scorzelli for Pope Paul VI, based on a traditional type from the Val Gardena region in northern Italy. In 1990, John Paul II was presented with a modified, lighter version.

However, we look in vain for a known story into which Cattelan's narrative fragment can be inserted. The artist himself does little to help. He is renowned for extreme and provocative imagery that he generally refuses to explain. He once encased the owner of the gallery that represents him in a huge, pink penis suit. The gallerist was even persuaded to wear the absurd costume for six weeks. This is the kind of stunt that gives the art world a bad name.

Cattelan stands in the tradition of Marcel Duchamp, the supreme maker of anti-art that the art world has canonized as art. In displaying everyday items in galleries, Duchamp challenged the definition of art and also exposed the art world to ridicule. Similarly, in Cattelan's piece, enigma and paradox prevail. We have to make of it what we will, aware that the joke might be at our expense.

Cattelan leaves some clues. The title, *La Nona Ora*, or The Ninth Hour, refers to the time of Christ's death on the cross. This representation of the death of Pope John Paul II might be an imitation of Christ's. In a typically elusive interview, Cattelan said, "I like the idea that someone is trying to save the Pope, like an upside-down miracle, coming not from the heavens but from earth". But he adds dismissively, "in the end it is only a piece of wax".

We may add gloss to his statement by saying that the death of a martyr involves human agency, followed by divine redemption, whereas Cattelan's Pope has been struck down by heavenly intervention and awaits earthly assistance. Our responses can range from seeing the image as moving and pious, evoking our sympathy with him as a modern martyr, to regarding it as shockingly blasphemous.

When the sculpture was shown in John Paul II's native Poland in 2000 at Warsaw's Zacheta gallery, shock prevailed. Two members of the Polish parliament tried to remove the meteorite and demanded the dismissal of the gallery's curator, Anda Rottenberg, whom they described as a "civil servant of Jewish origin". Rottenberg was eventually coerced into resigning.

My mind turns to the stone of the Kaaba in Mecca, Saudi Arabia, the focus of supreme devotion for Muslims, which is said to have been presented to Abraham by the Archangel Gabriel. It has been interpreted as a meteorite. Could Cattelan be alluding to the potential collapse of Christianity in the face of Islamic militancy? This would be inflammatory to both religions. However, it is the nature of art that the beholder completes the meaning of the artist's creation. Cattelan invites us to do so in extreme and contradictory terms.

Aware of the recent assaults on religion by scientific atheists, some people may even be tempted to see the felled Pope as an allegory of the conflict between extreme Darwinists and spiritual belief. Catellan's response might be that, although it is not actually 'wrong', this meaning is unintended. There is more to it. ■

Martin Kemp is research professor in the history of art at the University of Oxford, OX1 1PT, UK.

See Editorial, page 1143.

ESSAY

Beyond the notes

The way performers shape notes brings music to life. **Nicholas Cook** argues that measuring these subtle changes can help us appreciate and replicate the performer's art.

The art of musical performance lies largely in nuance — in making notes longer or shorter than they are written, or in shaping their dynamics, articulation or pitch. Performers don't generally have explicit theories of these things, as it's all done by ear. But these often subtle changes to the written score are responsible for a great deal of what makes music memorable, moving and meaningful — and they can be measured.

This combination of cultural meaning and measurability makes musical performance a productive example of the relationship between science and the humanities. When musicology came into being in the nineteenth century, it was modelled on philology, the study of ancient texts. For this reason, musicologists have tended to think of music as a form of writing. But much of what performers do, and what listeners respond to, falls between the notes as musicologists construe them. This is where science comes in.

Measurements cannot capture cultural values, but people listening to music respond to specific sounds. These sounds are amenable to scientific study, providing insights into the cultural values they embody. I focus on classical piano performance, but my claim is more general: to understand music as performance, we must use scientific and humanities approaches in tandem.

Stick to the plot

Two of the most important aspects of musical performance are the shaping of the tempo and the dynamics. Tempo shaping is the lengthening or shortening of notes or phrases and is measured by extracting beat durations from sound. Dynamics shaping is the patterning of loud and soft notes, to create one-off accents or waves of increase and decrease. It can be extracted as a continuously varying value or as a series of discrete values associated with individual notes.

Musicologists and psychologists have generally focused on how such data relate to the structure of music, drawing on traditional, notation-based analytical methods. For instance, reading by means of a kind of reverse engineering from the performance back to the composition, they have shown how performers use various combinations of speed change and dynamic accents to underline structural breaks or bring out important points.

Line graphs of tempo and dynamics are hard to relate to the music we hear, but over the past two years software has been developed that incorporates these graphs within a music visualization program so that they scroll past a cursor as one listens. Other limitations to this type of approach are less tractable. If performance is analysed in terms of the score-based structure, one is deaf to aspects of the performance that have nothing to do with what is written down. In effect, this assumes that the point of performance is to reproduce a meaning that is already there on the printed page, but any jazz or pop performance demonstrates what an inadequate approach that is.

There is a further, more subtle, problem. Try dancing to a Chopin mazurka and it soon becomes clear that concert evocations of dance music have much more extravagant shaping than music that is for actual dancing. A tempo graph would show this, and so says something about the music one experiences. Its shape, however, is the result of several distinct factors. To understand what is going on, we need to break the data down into their component parts. The question is what those parts might be.

Musical movement

In the early 1990s, Henkjan Honing and Peter Desain suggested that the shaping of both tempo and dynamics in classical music performance can be explained in terms of three main components: note-to-note shaping, the composer's 'pulse', and hierarchical phrase arching. The third of these refers to the way performers get faster and louder as they play into a phrase, and softer and quieter as they come out of it, giving the music a kind of breathing quality. It is often seen in nineteenth-century piano music, such as Chopin's. It is hierarchical in that such patterns can be found at multiple levels — such as 2, 4, 8 and even 16 bars. It is widely seen as part of what it means to play 'musically', that is to say expressively and meaningfully.

Musicologists tend to be suspicious of such generalizations. What is considered 'musical' has varied throughout history, as have practices of performance. My team at the AHRC Research Centre for the History and Analysis of Recorded Music (CHARM) in London recently analysed phrase arching in recordings of Chopin's *Mazurka* Op. 63 No. 3 going back to 1923. We measured how much shaping of phrases occurred through tempo and dynamics, and how far these variables were correlated. We found that, for this piece at least, both tempo and dynamic phrasing were present in the earliest recordings, but that they began to be closely coordinated with one another and with the composed phrasing only after the Second World War. Different performers achieved this coordination in various ways, but the effect was a streamlined style that was still expressive, albeit less personal and subjective than pre-war interpretations.

This finding shows how the nature of what is regarded as musicality has changed. What was assumed to be a general, perhaps hard-wired, quality turns out to be specific to a given time and place. Indeed, the very idea of 'expressive' performance, defined in terms of nuance, assumes that the purpose of music is to convey subjective feeling — an idea foreign to Japanese taiko drumming, for example. That is why these studies concentrate almost exclusively on Western classical music.

Mechanical musicians

It should be possible to apply a fully functioning model of expressive performance to a digital score that computers and synthesizers can read, in which every crotchet (US quarter note) is the same length.

D. PARKINS

This would result in an expressive and meaningful sound output that is mechanically generated but sounds human, reducing creativity to a set of rules.

Some programs do this on the basis of note-to-note shaping and composer's 'pulse'. Director Musices is a free, research-oriented program that uses a set of rules, for example that longer notes are louder than shorter ones, or that a run of ascending notes gets faster; there is also a simple phrase-arching function. These rules can be switched on or off, applied more or less strongly, or even inverted. The commercial program SuperConductor is based on the idea that for every major composer there is a characteristic signature (or pulse) for each beat in the bar, and allows you to 'sculpt' a file into an expressive performance. A new program currently in beta testing, Silbert MOR Expressive Performance, automatically generates human-like performances and is targeted at professionals such as the producers of TV commercials who want music without the trouble and expense of paying musicians or licensing recordings.

Such tools are a far cry from the 'humanize' functions of sequencing programs, which merely introduce random variation. And if, as the CHARM research suggests, performance styles can be modelled to specific times or places, variable settings could enable one to reproduce the style of particular pianists, effectively generating new recordings of pieces they never played. You might even fantasize about music being mixed in the same way as paint. Instead of buying recordings off the peg, like standard paint ranges, you could customize them: 50% Vladimir Horowitz, 45% Arturo Michelangeli and 5% Jean-Marc Luisada, say.

But a musical performance isn't a pot of paint. It is a human action carried out at a certain time and place, normally in the presence of others and marked by the contingencies of the occasion. The same applies to recordings, even when they owe more to studio manipulation than real-time performance. We still hear them as traces of events. Remove the communication from music and it rapidly becomes as pointless as it would be to spend time in the virtual world Second Life if there were no real people behind the avatars and speech bubbles.

"You might even fantasize about music being mixed in the same way as paint."

A search for meaning

Performance, then, is more than the communication of structural information about musical works. The very act of performance generates meaning, whether the musician is Madonna, Miles Davis or Glenn Gould.

In a 2002 concert performance of *Mazurka* Op. 63 No. 3 filmed at the Théâtre des Champs-Elysées in Paris, Russian pianist Grigory Sokolov performs virtuosity as much as he performs Chopin: his hands often fly up after a particularly telling note, providing an idiosyncratic balletic correlate to the sound. His performance makes perfect sense on CD, but seeing it adds further meaning. The striking quality of public display in his playing is redolent of the cavernous spaces of modern concert halls and the star quality of the international virtuoso. He enacts exceptionality.

Such evocative flourishes communicate cultural values that cannot be measured. Sokolov, like many Russian pianists, uses particularly strong phrase arching. His expressiveness is structurally generated, rather than primarily located at the note-to-note level as with pre-war pianists, so he is free to indulge in extravagant choreography without losing the musical thread.

Quantitative analysis reveals how phrase arching facilitates Sokolov's virtuosity. Without a systematic approach we would have much less idea about how these effects are created. It would be hard to quantify how Sokolov's style relates to that of other pianists.

Programs such as Director Musices or the CHARM model of phrase arching can be used to capture the general qualities of performance. The mark of their success is the extent to which they account for the variance in performance data. Such applications can also be used to study a particular performance, such as Sokolov playing Op. 63 No. 3. Here the interest lies in the pattern of discrepancies between the model and the performance. The focus is on the unique features, and the criterion of success is: how far the model guides the ear towards an awareness of these qualities, resulting in a process of engaged listening and critical interpretation. Used thus, deterministic models of performance expression do not undermine values of human creativity, but locate them more accurately.

Scientific measurement and cultural approaches to performance can be linked usefully. But this is a marriage of complementary approaches, rather than a convergence towards a unified discipline. ■

Nicholas Cook is director of the AHRC Research Centre for the History and Analysis of Recorded Music, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK. He is the author of *The Schenker Project: Culture, Race and Music Theory in Fin-de-siècle Vienna*.

For further reading see http://tinyurl.com/62zh2v. See other essays in the Science & Music series at www.nature.com/nature/focus/scienceandmusic.

ESSAY

The other beetle-hunter

Thanks to a fateful letter, the theory of evolution by natural selection was unveiled 150 years ago this week. **Andrew Berry** and **Janet Browne** celebrate the letter's writer, Alfred Russel Wallace.

One hundred years ago, to mark the 50th anniversary of the reading of the original papers by Charles Darwin and Alfred Russel Wallace on evolution by natural selection, the Linnean Society of London issued its first Darwin–Wallace awards to honour contributors to the study of evolution. Six of the seven 1908 recipients, including Francis Galton, Ernst Haeckel and Joseph Dalton Hooker, received silver medals. The only gold medal ever awarded went to Alfred Russel Wallace. At 85, he had five years to live and three books still to publish.

Wallace, who usually avoided academic ceremony, came to London from his home in Dorset for the occasion. His speech on "Why did so many of the greatest intellects fail, while Darwin and myself hit upon the solution of this problem" is vintage Wallace, a mixture of self-deprecation and insight. His conclusion? "In early life both Darwin and myself became ardent beetle-hunters."

Wallace went on to play down his role in the announcement of evolutionary theory. Indeed, in one account of the 1908 celebrations, his presence — and his speech — was entirely overlooked. The botanist Joseph Hooker was instead fêted as the "sole survivor of those immediately concerned".

It is too easy to see Wallace as the 'other man' of evolutionary theory, the one who served merely as a stimulus to Darwin. Worse, he is often remembered as a crank whose later embrace of spiritualism and socialism muddied his biological thinking.

In fact, he was a superb scientist, whose contributions to many aspects of evolutionary biology and biogeography remain influential. His conduct in the evolution business is exemplary. Despite rumblings from conspiracy theorists that Darwin cheated him, Wallace got exactly what he wanted: scientific recognition. Darwin too got what he wanted: precedence. And the book that reinforced that precedence will justly be celebrated next year as the foundation of modern biology.

Neither man expected the joint announcement of evolution by natural selection at the Linnean Society in 1858. Indeed it was not as self-sacrificing an arrangement as is often portrayed. And it exemplifies what scientists have always known — that the making of a new theory rarely occurs in isolation. Rather, it depends on the support of colleagues, social networks and interactions within the scientific community, as well as the power of the theory itself.

Humble beginnings

Wallace was born in 1823 into a middle-class family in decline. After a minimal education he became an assistant to his brother, a railway surveyor. Trekking around the English countryside, surveying-pole in hand, he became interested in natural history. After a downturn in surveying, Wallace spent a year as a schoolteacher in Leicester. Here, in 1844, he met Henry Walter Bates, a 19-year-old with great expertise in natural history, especially beetles. Wallace duly became an "ardent beetle-hunter". That same year, Robert Chambers anonymously published his controversial, flawed and widely read theory of evolution, *The Vestiges of the Natural History of Creation*, in which he proposed a universal "law of development". Wallace regarded this an "ingenious hypothesis".

Inspired by Darwin's and Alexander von Humboldt's published accounts of their journeys, Wallace and Bates headed to the Amazon in 1848. They funded their travels by selling exotic specimens to museums and collectors. The contrast with Darwin's voyage is striking. Being of considerable independent means, Darwin travelled in some style on the *Beagle* as the captain's paying guest. Wallace and Bates had to work for a living, depended on the hospitality and assistance of locals, and needed an agent in London to market their wares.

Wallace returned from Brazil in 1852 after four years of exploration, collection and privation. The trip ended in disaster: he lost nearly all his specimens, and almost his life, when his ship caught fire in the mid-Atlantic. With nothing to show for all his efforts, his hope of joining the scientific élite was cruelly derailed. In 1854, he set off for Southeast Asia to do it all over again.

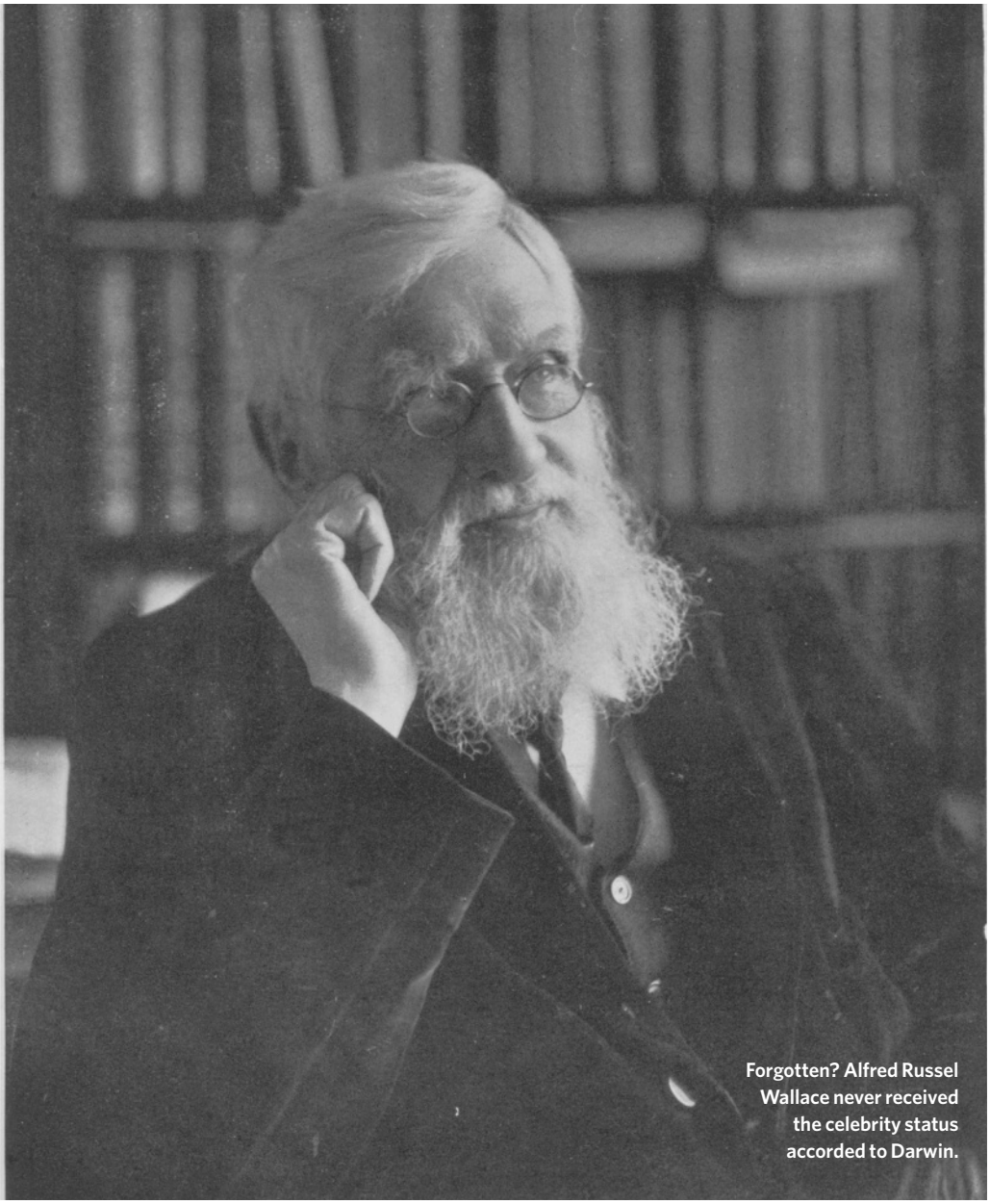
A year or so into these eastern travels, he was confident enough to write what he regarded as an evolutionary manifesto. "On the law which has regulated the introduction of new species" was published in 1855 in the *Annals and Magazine of Natural History*, a respected periodical read by both amateurs and professionals. Wallace pointed out that related species tend to occur together in both space and time — in the same geographical regions and in the same geological strata. The implication was clear to him: life consisted of a diversifying genealogical process. The paper was a major step towards the scientific status that Wallace craved, but it failed to create the stir he had hoped.

Around the start of 1856, geologist Charles Lyell told Darwin about Wallace's paper, warning Darwin that he might be scooped. Edward Blyth, an English naturalist in Calcutta, also wrote to Darwin: "Wallace has, I think, put the matter well; and according to his theory the various domestic races of animals have been fairly developed into species." In May 1856, not especially worried about Wallace, Darwin began to write the long-planned tome he expected to call 'Natural Selection'. He opened a correspondence with Wallace, noting that Lyell and Blyth had drawn his attention to the paper and sympathizing over the apparent lack of scientific reaction: "very few naturalists care for anything beyond the mere description of species". Better still, Darwin wrote that he agreed with Wallace's conclusions. Wallace was thrilled. Here was a direct connection to a major star of the scientific firmament.

Wallace's 'law' was still only half a theory of evolution. In February 1858, during a bout of malaria, he glimpsed the other half: the missing mechanism. Recalling the writings of the economist Thomas Malthus, Wallace suddenly recognized that better-adapted groups would gradually replace less well-adapted ones. He waited anxiously for his fever to end so he "might at once make notes for a paper on the subject", which he entitled "On the tendency of varieties to depart indefinitely from the original type". He then did a surprising thing. Rather than submitting the paper directly to a journal, he sent it to Darwin. No one else had shown such interest in his work.

A striking coincidence

In June 1858 (the exact date is unknown), Darwin opened and read a brilliantly incisive handwritten essay that repeated most of his own account of evolution by natural selection. Late in the evening of 18 June 1858, he wrote to Lyell: "I never saw a more striking coincidence… if Wallace had my MS sketch written out in 1842 he could not have made a better short abstract!" Some Wallace scholars suggest that Darwin may have received this letter several weeks earlier and used the intervening period to polish his own ideas in the light of Wallace's. But the documentary record attests to the gradual formulation of



Forgotten? Alfred Russel Wallace never received the celebrity status accorded to Darwin.

R. HAINES/MARY EVANS PICTURE LIBRARY

Darwin's theory over the previous 20 years. In particular, Darwin already had a clear understanding of evolutionary divergence, the main principle that some accuse him of taking from Wallace. Wallace was not telling Darwin anything he did not already know.

Publication was just as important to nineteenth-century science as it is now. Struggles over priority were fiery affairs that could make or break careers. Wallace's article was ready to be published — and as far as Darwin knew, it might already have been sent elsewhere for publication. As Lyell had predicted, he was forestalled. Gentlemanly honour required him to let Wallace take the credit. But Lyell and Hooker urged Darwin not to lose his claim as the originator of the theory. They suggested that there was room for manoeuvre.

These manoeuvrings have exercised historians ever since. Hooker and Lyell proposed a double announcement, so that priority would be shared. Despite his misgivings, Darwin agreed and sent them selections from his writings that explained his views and established chronological priority. Lyell and Hooker rushed these and Wallace's essay onto the programme of an extra meeting of the Linnean Society at the end of the season that was rescheduled because of the death of botanist Robert Brown, a former president of the society. Often described as a joint paper, it was rather two independent statements of the same idea. One hundred and fifty years ago this week, at a meeting on 1 July 1858, Lyell and Hooker communicated "On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection" to the Linnean Society.

Neither author was present. Darwin was wretched with grief over the death of his youngest child from scarlet fever two days earlier, and Wallace was seriously ill at Dorey (now named Manokwari) in New Guinea.

When Wallace heard about the fate of his essay, he immediately wrote to Darwin and the others to say that he thought the arrangements were completely satisfactory. To his mother he wrote: "I have received letters from Mr. Darwin and Dr. Hooker, two of the most eminent naturalists in England, which has highly gratified me. I sent Mr. Darwin an essay on a subject on which he is now writing a great work. He showed it to Dr. Hooker and Sir C. Lyell, who thought so highly of it that they immediately read it before the Linnean Society. This assures me the acquaintance and assistance of these eminent men on my return home…."

Wallace had made it. Like Darwin, although by a more arduous route, Wallace had gone from 'ardent beetle-hunter' to scientific luminary. This shared collecting spirit provided a link that lasted even when their intellectual paths began to diverge.

The papers were published in the Linnean Society's journal in August 1858, while Wallace was travelling to Ternate in the Moluccas. Darwin was by then working on what would become *Origin of Species*. Contrary to the usual story, several people recognized the likely impact of the Linnean Society papers: the American botanist Asa Gray, a close friend of Hooker and Darwin, immediately mentioned in print the value of evolutionary theory for explaining patterns of plant distribution; and a young ornithologist at the University of Cambridge, Alfred Newton, sat up all night to master their proposals. That said, Thomas Bell, president of the Linnean Society, guaranteed himself an unfortunate footnote in the history books by writing in his annual review of 1858: "The year which has passed has not, indeed, been marked by any of those striking discoveries which at once revolutionize, so to speak, the department of science on which they bear."

A new science

It took the *Origin of Species* to effect that revolution. One year later, with Darwin's book in his hands, Wallace was enthralled: "Mr. Darwin has given the world a new science," he wrote to his friend George Silk, adding that "his name should stand above that of every philosopher of ancient or modern times. The

force of admiration can no further go!!!"

So why has the name of one so prescient, and so generous, faded from popular view, while it still inspires those who find the modern infatuation with Darwin stultifying?

Exploring Wallace's role in the evolutionary story reveals a host of other figures who also deserve to be heard. Over the past twenty years, the Darwinian revolution has been shown to be neither a revolution as commonly understood nor solely due to Darwin. Many people proposed developmental schemes, some as famous as Jean-Baptiste Lamarck and Herbert Spencer, others relatively unknown but just as interesting. To remember Wallace is therefore to recognize that "evolution was in the air", and prompts one to wonder how Darwin's name rose so smoothly to the top.

The structure of science plays a part. First, the scientific community and the public tend to see science as a succession of advancing steps, each achieved by a named individual. In this view, precedence is everything: posterity ignores the second placed. Second, major changes in scientific theory are not just about the formulation of new ideas, but also depend on circulation and discussion. Shortly after Darwin's book was published, the word 'darwinism' began appearing in reviews and articles, and quickly came to denote an intellectual movement that also drew on the work of other figures, including Spencer, Chambers, Thomas Henry Huxley and Haeckel, as well as Wallace. Darwin's *Origin of Species*, and Darwin himself, became the flag to which many radical ideas rallied.

Perhaps Wallace contributed to his own eclipse too. For instance, he called one of his finest books *Darwinism*. Darwin's publishing strategy after *Origin of Species* was to consolidate, producing ever more evidence in support of the theory. Wallace, meanwhile, published on myriad topics, from the true identity of Shakespeare to the advisability of railway labour strikes. Darwin's politics, although strongly felt, had few public airings. Wallace, in contrast, was an outspoken socialist, the campaigning president of the Land Nationalisation Society, which insisted that private ownership of land was the root of all social iniquity. Attracted to radical issues, he became a spiritualist, believed in phrenology as "the true science of mind", and was a leading opponent of smallpox vaccination. This undermined his credibility with many scientists. Some defenders of Wallace consider him a victim of the Victorian class system, but his problems stem from more than a humble background. After all, Huxley, Darwin's most prominent advocate, was born above a butcher's shop yet became the leading spokesman for British science.

"Wallace compared Darwin to a great military general and likened himself to a guerrilla, useful for a skirmish."

Step by step, Darwin's star brightened as Wallace's faded. By the time Darwin died, he was held to be "first among the scientific men of England", as the socialist writer Edward Aveling put it. Darwin's name was inextricably linked with the idea of evolution and with broader shifts in public opinion that swept through the nineteenth century. Wallace never acquired Darwin's celebrity status. Unlike Darwin, he was not buried in Westminster Abbey, although a wall medallion was unveiled there in 1915, two years after his death. None of his houses became a museum. Images of Wallace did not appear in any of the satires or cartoons of evolution. Nor did Wallace have the evocative connection that Darwin did to the Galapagos Islands. His manuscripts were not published, and his library was not preserved.

At the start of the twenty-first century, Darwin could hardly be more prominent. His name is invoked in every modern discussion of evolution. He stares out from websites both for and against evolutionary theory. Books, stamps, exhibitions, conferences, festivals and artistic works abound. A portrait of Darwin was commissioned in 1881 by the Linnean Society from the artist John Collier, and copied for the National Portrait Gallery, the Royal Society and the Darwin Museum at Down House. By contrast, the portrait of Wallace that hangs in the Linnean Society was not painted until 1998.

NAT. HIST. MUS., LONDON

Two of Wallace's notebooks.

Wallace modestly endorsed these differences. In a letter in 1869, he compared Darwin to a great military general who kept sight of every campaign detail, and likened himself to a guerrilla, useful for a skirmish. "I feel truly thankful that Darwin had been studying the subject so many years before me, and that I was not left to attempt and to fail, in the great work he has so admirably performed."

As for the events of 150 years ago, Erasmus Darwin, Charles's older brother, encapsulated Wallace's magnanimity when he wrote in 1871 to Charles's daughter Henrietta: "in future histories of science the Wallace–Darwin episode will form one of the few bright points among rival claimants." ■

Andrew Berry is lecturer on biology at Harvard University, Biology Laboratories, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA, and is editor of an annotated anthology of Wallace's writings, *Infinite Tropics*.
Janet Browne is Aramont professor of the history of science, Department of the History of Science, Harvard University, Science Center 371, Cambridge, Massachusetts 02138, USA. She is the author of a two-volume biography of Darwin and of *Darwin's "Origin of Species": A Biography*.

See http://tinyurl.com/5beghd for further reading.

NEWS & VIEWS

PLANETARY SCIENCE

Forming the martian great divide

Walter S. Kiefer

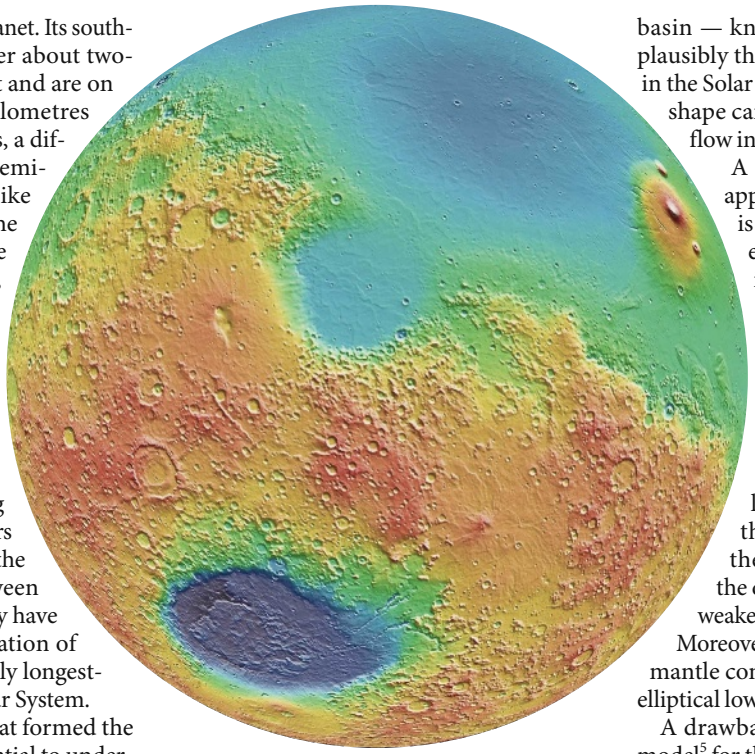
Early in its history, Mars suffered a convulsion that left a lasting geological and topographical scar. The latest work adds to evidence that the cause was external — a massive impact.

Mars is a divided planet. Its southern highlands cover about two-thirds of the planet and are on average about 4 kilometres higher than the northern plains, a difference that is known as the hemispheric dichotomy[1] (Fig. 1). Like an ice cube floating in water, the high topography is held up by the buoyancy of thicker crust (Fig. 2, overleaf) — the crust is about 25 km thicker in the highlands than in the lowlands[2]. On the basis of the number of impact craters in both the highlands and lowlands, the dichotomy is thought to have formed more than 4 billion years ago, during the first few hundred million years of martian history[3]. Moreover, the location of the boundary between the highlands and lowlands may have controlled the subsequent location of Tharsis[4], the largest and possibly longest-lived volcanic region in the Solar System.

Unravelling the processes that formed the hemispheric dichotomy is essential to understanding the earliest history of Mars. Previous explanations have invoked either the impact of a large asteroid or comet[5], or large-scale convective circulation in the martian mantle[6]. But observations by spacecraft have not yet permitted a clear choice between these possibilities.

Three papers in this issue[7–9] provide insight into this problem, and collectively strengthen the plausibility of the giant-impact model. One of the difficulties in testing the different hypotheses is that 30% of the boundary between highlands and lowlands is masked by later Tharsis volcanism. Andrews-Hanna *et al.*[7] (page 1212) used gravity observations to subtract the contribution of Tharsis volcanism from the crust and thus estimated the crustal signature of the dichotomy boundary in this part of Mars.

The key to their approach is that, because Mars has cooled with time, its elastic lithosphere (the strong, outermost layer of the planet) was thicker when the Tharsis volcanoes formed than when the hemispheric dichotomy formed. Assuming that the lithosphere was at least 100 km thick by the time the bulk of Tharsis volcanism occurred, Andrews-Hanna *et al.* show that the northern lowlands have an elliptical shape that is roughly 10,600 km by 8,500 km across. The statistical uncertainty in the basin dimensions is a few hundred kilometres. The authors conclude that the elliptical basin — known as the Borealis basin — is plausibly the signature of the largest impact in the Solar System. They also argue that this shape cannot be produced by convective flow in the mantle.



NASA

Figure 1 | The martian hemispheric dichotomy. This image of Mars's eastern hemisphere provides a vivid depiction of the dichotomy, with the southern highlands appearing in red–yellow and the northern lowlands in blue. The hemispheric dichotomy may be due to the impact of a massive asteroid early in martian history[7–9], producing a proposed ~10,000-km-wide impact structure called the Borealis basin as the northern lowlands. The Hellas basin (lower left) is also an impact structure, but at 2,300 km across is much smaller. Another of Mars's geological structures, the Tharsis volcanic region, is on the opposite side of the planet to that shown here.

A potential problem with the approach of Andrews-Hanna *et al.*[7] is its sensitivity to the thickness of elastic lithosphere. Although some independent evidence[10] supports the assumption of a thickness of 100 km or more, other observations[11] suggest that the lithosphere was less than 20 km thick during the Noachian period, about 3.8 billion years ago, when Tharsis began forming. If the lithosphere was less than 50 km thick during Tharsis formation, the elliptical shape calculated for the dichotomy lowlands is degraded[7], weakening the case for an impact model. Moreover, no calculations have shown that mantle convection is unable to produce an elliptical lowland basin.

A drawback of the original giant-impact model[5] for the dichotomy is that the resulting basin was assumed to be circular, as are the vast majority of small impact craters in the Solar System. By contrast, the northern lowlands are clearly elongated, with an eccentricity of ~1.2 (ref. 7). Three-dimensional hydrodynamic impact simulations by Marinova *et al.*[8] (page 1216) now show that the original reasoning was incorrect. Small impact craters are essentially formed on a flat surface, and thus are not sensitive to the spherical shape of the planet. But in an impact large enough to form the hemispheric dichotomy, one must consider the interaction of the impactor with the spherical planet. Marinova *et al.* show that, for planet-scale impacts, an oblique impact angle of 30–60° from the horizontal can produce the observed basin eccentricity. This range of impact angles is the statistically most likely range, and the authors' inferred impact velocity of 6–10 km s^{-1} is reasonable for a body intersecting the orbit of Mars. If the impactor was a rocky or metallic asteroid, it would have had to have been between 1,600 km and 2,700 km

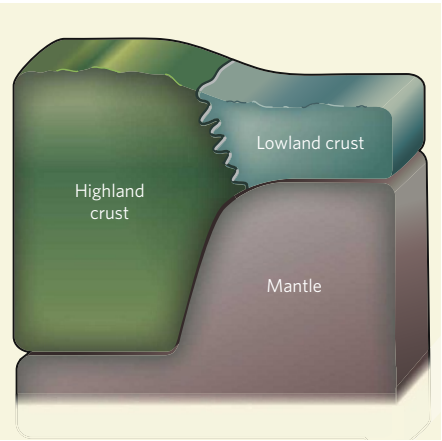


Figure 2 | Crust with a difference. The southern highlands (left) consist of relatively thick crust and high-standing topography. The northern lowlands (right) consist of thinner crust and lower topography. In the impact model for dichotomy formation, which is supported by the new work[7–9], the crust also differs in composition. According to this model, the highland crust formed early in martian history, and consists primarily of basalt rock. The lowland crust formed by shock melting of the mantle during the impact event, and so is both younger and different in composition compared with the highland crust.

in diameter to produce the observed basin size. For comparison, Mars itself is 6,780 km in diameter.

A limitation of the three-dimensional simulations done by Marinova *et al.*[8] is that the vertical resolution is 118 km. This is about twice as large as the estimated average crustal thickness[2], and thus details of crustal excavation and the resulting basin shape are not well resolved in these simulations. Nimmo *et al.*[9] (page 1220) have developed a complementary set of numerical models with a vertical resolution of 25 km in the crust and upper mantle. This provides a much better resolution of the behaviour of the crust during an impact, although the models are two dimensional and thus limited to vertical impacts.

In Nimmo and colleagues' models, almost all of the pre-existing crust is excavated from the centre of the impact basin, and a new, thinner crust is generated by shock melting of the martian mantle. The inferred impact energy is a factor of 3–5 lower than in the simulations of Marinova *et al.*[8]. This disparity is probably due mostly to the difference between vertical and oblique impacts. Nevertheless, the overall concordance of the two sets of models suggests that we now have a reasonable first-order understanding of the dynamics of the proposed impact event.

These three papers[7–9] strengthen the plausibility of the impact model for the hemispheric dichotomy on Mars, but they do not rule out mantle convection as the primary cause. One way to further test the impact model would be to gain a better understanding of the topographical variability along the proposed impact basin rim. In some places, the transition between highland and lowland occurs over just a few hundred kilometres, whereas in others a gradual change in elevation occurs over several thousand kilometres. The two boundary types also have distinct patterns in the gravity data. These differences have been proposed to be consequences of post-impact flow in the crust and mantle[7,12]. Development of quantitative models for these processes would be a great help in understanding the overall geological evolution of the proposed Borealis basin.

Another test relates to crustal composition in the highlands and lowlands. In the giant-impact model, most of the lowland crust formed by shock melting of the mantle[9] (Fig. 2). Extraction of the highland crust from the mantle modified the mantle's initial chemical composition. Impact-shock melting of the modified mantle is likely to have produced a lowland crust that differs in composition from that of the highlands. Some differences in composition between highlands and lowlands have been proposed on the basis of spectroscopy measurements made by orbiting spacecraft[13]. Impact-cratering statistics also suggest that the lowlands are slightly younger than the highlands[3]. Both observations are consistent with the impact model.

A more precise test of possible differences between highlands and lowlands would involve comparative analysis of bedrock composition in the two regions by future robotic landers. Landers could also measure how seismic-wave velocities vary with depth, which would provide a more comprehensive test of possible variations in composition throughout the entire thickness of the crust. Seismic measurements of crustal structure could also directly probe the proposed location of the dichotomy boundary in southern Tharsis.

Meantime, the main cause of Mars's great divide must remain a story without an ending. But appropriately enough, given the hundredth anniversary of the Tunguska event in Siberia celebrated elsewhere in this issue, these three sets of authors[7–9] have shortened the odds that it was produced when a huge impactor collided with a young Mars in what would have been an event of awesome proportions. ■

Walter S. Kiefer is at the Lunar and Planetary Institute, 3600 Bay Area Boulevard, Houston, Texas 77058, USA.
e-mail: kiefer@lpi.usra.edu

1. Watters, T. R., McGovern, P. J. & Irwin, R. P. *Annu. Rev. Earth Planet. Sci.* **35,** 621–652 (2007).
2. Neumann, G. A. *et al. J. Geophys. Res.* **109,** E08002, doi:10.1029/2004JE002262 (2004).
3. Frey, H. V. *J. Geophys. Res.* **111,** E08S91, doi:10.1029/2005JE002449 (2006).
4. Zhong, S. *Lunar Planet. Sci. Conf.* **39,** abstr. 1528 (2008).
5. Wilhelms, D. E. & Squyres, S. W. *Nature* **309,** 138–140 (1984).
6. Roberts, J. H. & Zhong, S. *J. Geophys. Res.* **111,** E06013, doi:10.1029/2005JE002668 (2006).
7. Andrews-Hanna, J. C., Zuber, M. T. & Banerdt, W. B. *Nature* **453,** 1212–1215 (2008).
8. Marinova, M. M., Aharonson, O. & Asphaug, E. *Nature* **453,** 1216–1219 (2008).
9. Nimmo, F., Hart, S. D., Korycansky, D. G. & Agnor, C. B. *Nature* **453,** 1220–1223 (2008).
10. Phillips, R. J. *et al. Science* **291,** 2587–2591 (2001).
11. McGovern, P. J. *et al. J. Geophys. Res.* **109,** E07007, doi:10.1029/2004JE002286 (2004).
12. Kiefer, W. S. *Geophys. Res. Lett.* **32,** L22201, doi:10.1029/2005GL024260 (2005).
13. Karunatillake, S. *et al. J. Geophys. Res.* **111,** E03S05, doi:10.1029/2006JE002675 (2006).

See Editorial, page 1143.

BEHAVIOURAL NEUROSCIENCE

Out of sight, but not out of mind

Seth M. Tomchik and Ronald L. Davis

Flies are cleverer than previously thought. They can remember their original destination even if distracted en route by another landmark. This behaviour depends on a specific group of neurons.

You are walking down a street to meet a friend at the end of it. You are early; so to kill time, you go into a café. After this brief detour, you continue on your way to meet your friend. While in the café, your original destination was out of sight, yet your brain held that goal in memory. Many such distractions occur during our daily tasks, yet in most cases we can remember and complete the original task. On page 1244 of this issue, Neuser *et al.*[1] demonstrate that the fruitfly *Drosophila melanogaster* can perform a similar task, and elucidate some of the mechanisms that the fly brain uses for this type of memory.

To determine whether flies can retain memory of a landmark when presented with a distraction, the authors used a modified version of Buridan's paradigm[2]. For the original paradigm, each fly is placed in a circular arena with two opposing black, vertical stripes on its walls. Normally, a fly will pace back and forth between the two stripes (Fig. 1a). But if the stripes are removed as the fly crosses the midline of the arena, it will maintain its heading, which suggests that it remembers the location of the stripes or its own trajectory[3].

Neuser *et al.*[1] devised a variation called the detour test. They removed the original target stripes as the fly was halfway through its second crossing, and immediately presented a distracter stripe on the arena wall either to the left or the right of the insect (Fig. 1b). Once the

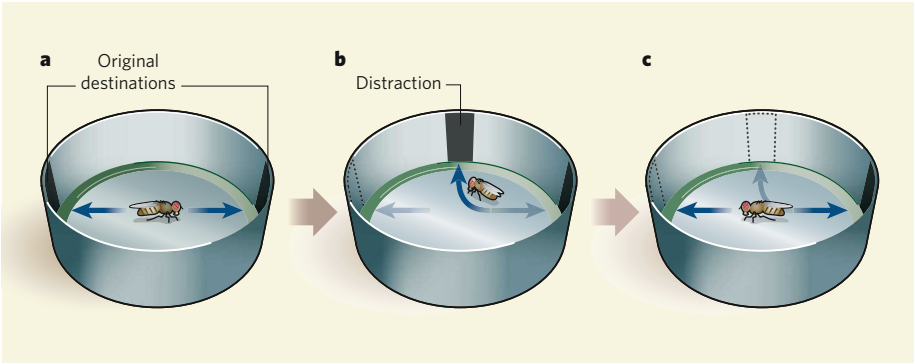


Figure 1 | Undeterred by the detour. **a**, A fly will walk back and forth between two opposing black stripes in a circular arena even if, midway, the target stripe disappears. Neuser *et al.*[1] find that if (**b**) a fly is distracted midway by a new black stripe to its side, the insect still remembers the position of the original stripe, and (**c**) when the distracter stripe subsequently disappears, resumes walking along its original course. The authors find that this behaviour depends on ring neurons of the ellipsoid body in the fly brain.

fly turned and headed towards the distracter, the investigators made that stripe disappear too. In most cases, the fly then turned back and reassumed its initial target orientation, suggesting that it still remembered the location of the original stripe (Fig. 1c). The distracted flies' memory of the location of the original stripe lasted at least four seconds and might last longer, as the authors did not explore the maximum time for which the original direction was remembered.

What is this interesting type of memory that might be used for navigation? Performance in the detour paradigm seems to be innate and hard-wired. Flies have a natural tendency to move towards arena landmarks[2]: they remember the position of a stripe without being trained to remember it or to associate it with a reward or punishment; and several practice laps in the detour paradigm fail to improve the flies' accuracy in reorienting themselves towards the original target after they are distracted. Neuser *et al.*[1] describe this memory type broadly as 'spatial orientation memory'. But do the flies remember the location of the original stripe by forming a neural representation of this external spatial cue, or by remembering their own movements towards the distracter relative to their original orientation?

Perhaps both. There is extensive evidence[4] that specific types of neuron guide horizontal navigation in vertebrates. One well-studied type of neuron is the head-direction cell. These neurons are present in many brain areas, and become active only when an animal's head points in a specific direction, irrespective of the animal's location or its behaviour at the time. Initially, the optimal firing properties of these neurons can be guided by external landmarks, but they are maintained when the landmarks are no longer visible. Neuser and colleagues' detour paradigm for *Drosophila* might tap into a memory that is partly mediated by neurons akin to the head-direction cells of vertebrates: that is, the memory forms as a result of association between directional neurons' activity and either a neural representation of the original visual target or cues derived from the animal's own movement.

Different types of memory are often stored in specific brain areas[5]. To localize the regions of the fly brain that are crucial for performance in the detour paradigm, the authors tested several lines of mutant flies that have structural defects in the central complex — an area of the brain involved in coordinating movement in insects[6]. Mutants with defects in a part of the central complex called the ellipsoid body performed poorly in the detour paradigm, suggesting that this brain structure is necessary for the task. The researchers verified the requirement for the ellipsoid body by genetically silencing subsets of its neurons in flies and then testing their performance. This revealed that ring neurons within the ellipsoid body are specifically required for performance in the detour test. As these neurons release the inhibitory neurotransmitter GABA (γ-aminobutyric acid), they are thought to inhibit downstream neurons.

Do the *Drosophila* ring neurons participate in spatial orientation memory by monitoring direction — like the head-direction neurons of vertebrates — or by forming a neural representation of the original visual target? Again, perhaps both, but there are also other possibilities. For instance, ring neurons might not be part of the neural system that forms the memory, but might instead have some ancillary role in the circuitry required for task performance. Interestingly, the mushroom body — a part of the fly brain necessary for certain tasks such as forming memories about odours and retrieving learned associations about visual cues independently of the context[5,7–9] — is not necessary for performance in the detour test.

Having discovered that ring neurons mediate spatial orientation memory, Neuser *et al.* embarked on identifying molecular signalling pathways that are involved in the detour task. The signalling molecule cyclic AMP was a logical candidate, because it is essential for olfactory learning in fruitflies[5]. The authors tested *dunce*, a cAMP-pathway mutant, in the detour task and found that, surprisingly, these

50 YEARS AGO

'Whistlers' are being heard consistently at Scott Base, the New Zealand International Geophysical Year Antarctic Station, McMurdo Sound. From the time observations commenced on April 15 until the time of writing considerable activity has been observed, including 'bonks', 'tweeks', short and long 'whistlers', whistler trains, and periods of strong 'sferics'. No dawn chorus has yet been observed... It is believed that this is the first time that whistlers have been heard in such a high geomagnetic latitude. The whistlers appear to have dispersion characteristics similar to those heard in lower geomagnetic latitudes. However, the characteristics cannot be truly determined until tape recordings of them are sent to New Zealand for spectrographic analysis, at the end of the Antarctic winter.

From *Nature* 28 June 1958.

100 YEARS AGO

"The rings of Saturn" — In a note published as Bulletin No. 32 of the Lowell Observatory, Prof. Lowell develops rather more fully the idea that the appendages B and C of Saturn are not flat rings, but tores. He arrives at this conclusion, by two independent methods, from a discussion of the phenomena observed at Arizona during November and December last. In the first place, a black core was observed running medially through the length of the shadowy band which then encircled the planet. This core...is presumed to be the black shadow of the plane ring A bordered by the particles of the rings B and C scattered above and below the plane of A. That is to say, the rings B and C differ from A in being tores and not flat rings... The assumed heaping up of the particles, as indicated by the agglomerations [seen at many observatories], is in accordance with gravitational laws. Furthermore, it is shown from the observational results that the inevitable disintegration of the rings is in the process of taking place.

From *Nature* 25 June 1908.

50 & 100 YEARS AGO

insects performed as well as normal flies. By contrast, another mutant impaired in olfactory learning, *ignorant*[10], did not perform well. This mutant is defective in the gene encoding the protein-kinase enzyme S6KII. The authors find that expressing S6KII specifically in the ring neurons of *ignorant* flies restores their performance in the detour test. S6KII might affect some forms of memory through the signalling pathway mediated by another protein kinase, MAPK, because S6KII is a downstream effector of MAPK (refs 10, 11).

Neuser and colleagues' work[1] offers insight into a newly discovered memory system in *Drosophila* that seems to be involved in choosing the direction of walk. Their findings add to the growing body of data suggesting that fruitflies are capable of extremely complex behaviours. Together, these observations indicate that flies can initiate goal-directed behaviour, remember the goal despite a distraction, and re-initiate and execute the behaviour necessary to reach the goal on removal of the distracter. Several questions remain, including how the ring neurons specifically participate in this memory system, whether they are part of a *Drosophila* equivalent of vertebrates' head-directional circuits, and what the other components of the memory circuit are. ■

Seth M. Tomchik is in the Department of Molecular and Cellular Biology, and Ronald L. Davis is in the Departments of Molecular and Cellular Biology, and of Psychiatry and Behavioural Sciences, Baylor College of Medicine, Houston, Texas 77030, USA.
e-mail: rdavis@bcm.tmc.edu

1. Neuser, K., Triphan, T., Mronz, M., Poeck, B. & Strauss, R. *Nature* **453,** 1244–1247 (2008).
2. Gotz, K. G. in *Development and Neurobiology of* Drosophila (eds Siddiqi, O., Babu, P., Hall, L. M. & Hall, J. C.) 391–407 (Plenum, 1980).
3. Strauss, R. & Pichler, J. *J. Comp. Physiol. A* **182,** 411–423 (1998).
4. Taube, J. S. *Annu. Rev. Neurosci.* **30,** 181–207 (2007).
5. Davis, R. L. *Annu. Rev. Neurosci.* **28,** 275–302 (2005).
6. Strauss, R. & Heisenberg, M. *J. Neurosci.* **13,** 1852–1861 (1993).
7. Zars, T. *Curr. Opin. Neurobiol.* **10,** 790–795 (2000).
8. Liu, L., Wolf, R., Ernst, R. & Heisenberg, M. *Nature* **400,** 753–756 (1999).
9. Mehren, J. E., Ejima, A. & Griffith, L. C. *Curr. Opin. Neurobiol.* **14,** 745–750 (2004).
10. Putz, G., Bertolucci, F., Raabe, T., Zars, T. & Heisenberg, M. *J. Neurosci.* **24,** 9745–9751 (2004).
11. Xing, J., Ginty, D. D. & Greenberg, M. E. *Science* **273,** 959–963 (1996).

DRUG DISCOVERY

A lifeline for suffocating tissues

Massimiliano Mazzone and Peter Carmeliet

When a blood vessel becomes blocked, the ideal treatment would be a drug that induces new vessel formation in the damaged tissue, without affecting healthy tissues. With the chemical nitrite, we might be on to a winner.

Ischaemia occurs, for instance, when a blood vessel becomes occluded by a clot. It affects hundreds of millions of people worldwide, and is often life-threatening. So there is an urgent need for molecular factors that could stimulate new vessel growth (angiogenesis) and so promote revascularization of ischaemic tissues. Progress so far has not been particularly noteworthy, and one major problem is that some potential angiogenic factors promote vascularization in healthy tissues as well as in the ischaemic tissue, causing undesirable side effects. Writing in *Proceedings of the National Academy of Sciences*, Kumar *et al.*[1] revive hope that a drug might one day become available. They report that systemic administration of nitrite (NO_2^-), which is reduced to nitric oxide (NO), for protracted periods restores blood flow in ischaemic tissue without stimulating angiogenesis in healthy tissues.

The idea of therapeutic angiogenesis is not new. Over the past decade, considerable efforts have gone into assessing whether the administration of angiogenic factors, such as growth factors, signalling molecules and gene transcription factors, would improve revascularization. But despite initial promise, these molecules have mostly failed in clinical trials, chilling the enthusiasm for the feasibility of therapeutic angiogenesis. Reasons for the disappointing clinical results include technical difficulties in efficiently administering the angiogenic factor locally to the ischaemic tissue over sufficiently long periods of time; the short metabolic half-life of these factors; and the requirement for multiple angiogenic factors to build a mature, stable, functional vasculature[2,3]. Moreover, the potential risk of side effects, such as deregulation of blood pressure and stimulation of dormant tumour growth, has precluded systemic administration of angiogenic factors, especially those that act indiscriminately on vessels in healthy and ischaemic tissues.

A large body of evidence[4,5] implicates NO — which was initially discovered as a factor that relaxes blood vessels — in stimulating angiogenesis. This signalling molecule increases the expression of various angiogenic factors, including VEGF, which, together with other mediators, increases NO levels through positive feedback. As well as stimulating the growth of nascent and immature vessels consisting only of fragile endothelial cells, NO recruits perivascular mural cells, which stabilize vessels, allowing them to become fully functioning conduits. Moreover, NO improves blood perfusion by inducing vessel dilation and possibly by enhancing the formation of collateral vessels, which supply the bulk of the flow to ischaemic tissues.

Another attractive property of NO is that it can protect tissues against ischaemic damage by slowing down cellular respiration. It does this at least in part through the nitrosylation of complex I proteins in the electron-transfer chain. This reduces the production of toxic reactive-oxygen species, which often occurs under ischaemic conditions[6]. What's more, at high NO levels, which may occur in inflamed tissues, nitrosylation inactivates oxygen sensors of the PHD family of proteins[7]. Inhibition of the PHD1 sensor protects tissues against ischaemic damage by reprogramming the cell's metabolism and reducing oxidative stress[8]. Both of these nitrosylating activities of NO therefore induce 'ischaemic tolerance'. So it is not surprising that NO has been considered a candidate for the revascularization and protection of ischaemic tissues.

But the therapeutic potential of NO is both context dependent and dose dependent. Despite its beneficial actions, it could have toxic effects at high concentrations. Indeed,

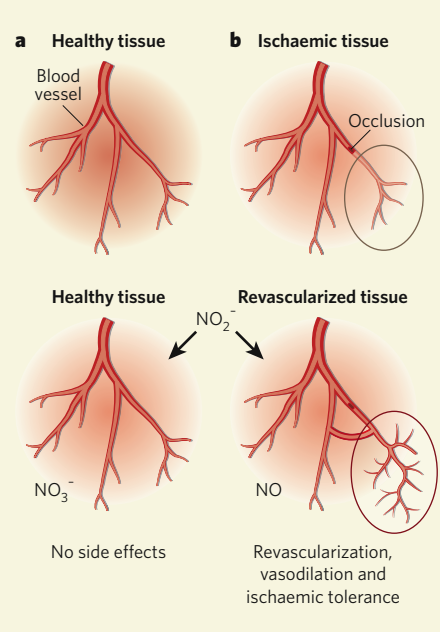


Figure 1 | Targeted effect. Kumar *et al.*[1] find that, when administered systemically in mice, (**a**) nitrite (NO_2^-) is converted to nitrate (NO_3^-) in tissues that contain normal perfused blood vessels. **b**, In ischaemic tissues, by contrast, nitrite is selectively converted to nitric oxide (NO), where it stimulates revascularization, vasodilation and ischaemic tolerance.

in addition to enhancing programmed cell death and decreasing cell proliferation, NO inactivates the oxygen carrier haemoglobin and inhibits the cytochrome *c* oxidase enzyme, thus impairing cellular respiration. Moreover, under certain conditions, it stimulates oxygen sensors and promotes the formation of reactive-oxygen species. Finally, at least when given systemically as an organic nitrate compound, NO acts indiscriminately in diseased and healthy tissues.

The duality of NO activity could explain why, paradoxically, both molecules that produce NO and inhibitors of the enzyme NO synthase, which catalyses NO production, protect cells against ischaemic injury[5]. And it raises questions about the 'safe window' of NO levels for therapy. So why did Kumar *et al.*[1] consider using nitrite as an angiogenic factor? In their study, the authors took into account the fact[5,9] that nitrite is reduced to NO under ischaemic conditions, whereas in well-oxygenated tissues it is oxidized to apparently harmless nitrate (NO_3^-; Fig. 1). In other words, nitrite acts as a site-selective 'pro-drug' by preferentially generating NO in ischaemic tissues, where it stimulates revascularization and cell protection, while avoiding potentially harmful NO generation in healthy tissues.

Such a safety profile would overcome the difficulty of administering an angiogenic drug locally at the ischaemic site, and instead might allow systemic — even oral — delivery over prolonged periods. Kumar *et al.* show that, in mice, sodium nitrite promotes revascularization of ischaemic tissues by stimulating the formation of mature, perfused vessels, whereas treatment with a NO scavenger abolishes this effect, thus supporting the idea that conversion of nitrite to NO underlies its beneficial effect in ischaemia.

Nonetheless, as is often the case, intriguing questions remain. Reactive-oxygen species that form during ischaemia interact with nitrite to generate highly reactive peroxynitrite, which can damage DNA and proteins[4,9]. Will such damage caused by oxidative stress occur with long-term nitrite treatment? Also, some of the beneficial effects of NO are mediated by its acute vasodilatory effects. Will the benefits of nitrite therapy persist after its withdrawal?

Kumar *et al.* suggest that nitrite accumulation in ischaemic muscle stimulates revascularization in a NO-dependent manner three days after the start of ischaemia; but they also observe that products of NO metabolism are detectable only after seven days. These puzzling observations remain to be reconciled. Moreover, the extent to which the protection nitrite offers against ischaemic tissue damage relates to ischaemic tolerance rather than revascularization remains to be explored.

Another recent study[10] documents cardioprotective effects of nitrite just one day after reperfusion of ischaemic heart tissue — too short a time-frame for angiogenesis to occur after permanent blockage of an artery. If it is ischaemic tolerance that precedes revascularization, how rapidly can nitrite induce it? Also, unlike chronic ischaemia, a severe acute ischaemic event is characterized by rapid cell death, and nitrite-mediated ischaemic tolerance might not suffice as a beneficial adaptation. So, will starting treatment after the acute event still be effective?

Finally, the dual activities of NO warrant a careful choice of the nitrite dose administered, ensuring that the optimal therapeutic dose of this drug won't cause a drop in systemic blood pressure. NO inhalation could be a promising delivery route that might induce less fluctuation in blood pressure. Furthermore, this delivery route selectively increases NO in ischaemic tissues, with cytoprotective results[11]. Kumar and colleagues' promising findings[1] certainly warrant further study of the therapeutic potential of this molecule. ■

Massimiliano Mazzone and Peter Carmeliet are in the Vesalius Research Center, University of Leuven, Flanders, Institute for Biotechnology (VIB), B-3000 Leuven, Belgium.
e-mail: peter.carmeliet@med.kuleuven.be

1. Kumar, D. *et al. Proc. Natl Acad. Sci. USA* **105,** 7540–7545 (2008).
2. Yla-Herttuala, S., Markkanen, J. E. & Rissanen, T. T. *Trends Cardiovasc. Med.* **14,** 295–300 (2004).
3. Tirziu, D. & Simons, M. *Angiogenesis* **8,** 241–251 (2005).
4. Luque Contreras, D., Vargas Robles, H., Romo, E., Rios, A. & Escalante, B. *Pharmacol. Ther.* **112,** 553–563 (2006).
5. Moncada, S. & Higgs, A. *N. Engl. J. Med.* **329,** 2002–2012 (1993).
6. Shiva, S. *et al. J. Exp. Med.* **204,** 2089–2102 (2007).
7. Metzen, E. *et al. Mol. Biol. Cell* **14,** 3470–3481 (2003).
8. Aragones, J. *et al. Nature Genet.* **40,** 170–180 (2008).
9. Dezfulian, C., Raat, N., Shiva, S. & Gladwin, M. T. *Cardiovasc. Res.* **75,** 327–338 (2007).
10. Gonzalez, F. M. *et al. Circulation* **117,** 2986–2994 (2008).
11. Liu, X. *et al. J. Am. Coll. Cardiol.* **50,** 808–817 (2007).

ATMOSPHERIC CHEMISTRY

Sun, sea and ozone destruction

Roland von Glasow

Halogens are known to decrease the levels of stratospheric ozone. The latest measurements show that something similar occurs in the lower atmosphere over tropical oceans — and probably above most other oceans, too.

Ozone is a fascinating atmospheric gas, with different roles depending on where it is located. In the troposphere (the lowest part of the atmosphere that extends about 10 kilometres above Earth's surface) it is a greenhouse gas, and can be harmful to plants and animals. But in the stratospheric ozone layer (at an altitude of around 25 km) it absorbs damaging ultraviolet radiation from the Sun. Ozone is also the main precursor of the hydroxyl radical — a highly reactive molecule that 'cleanses' the air of pollutants, and also helps remove methane, a potent greenhouse gas.

This cleansing of the atmosphere is especially pronounced in the tropics, where about 75% of global methane oxidation occurs. The processes that control ozone levels in the tropical troposphere are therefore of great interest. On page 1232 of this issue, Read *et al.*[1] provide compelling evidence that a crucial destruction pathway for ozone in the lower troposphere has so far been ignored. They present the first long-term study of halogen oxide and ozone concentrations in the marine boundary layer, the lowest kilometre of the atmosphere above the ocean. Their results from the Cape Verde islands in the tropical Atlantic provide the first confirmation that halogens have a substantial impact on regional, and possibly even global, ozone levels.

Read *et al.* are not the first to take measurements of halogen oxide concentrations in the boundary layer. For example, bromine oxide (BrO) levels have been measured in polar regions[2], and iodine oxide (IO) has been measured at several coastlines, including the west coast of Ireland[3]. These previously observed concentrations were often higher than those now reported at Cape Verde[1]. So what is special about Read and colleagues' study?

The main difference is that the new measurements were made in the open ocean, where they cannot be influenced by local features that produce halogens, such as the coastal kelp beds of Ireland. Also of note is that the halogen oxides at Cape Verde were present year-round, unlike in polar regions, where high BrO concentrations persist for only a few months. Most importantly, Read and colleagues' measurements[1] might be typical of conditions that prevail over the open oceans, which cover 70% of Earth's surface.

The halogen oxides BrO and IO are involved in efficient catalytic reaction cycles in which halogens destroy ozone (Fig. 1). In an effort to quantify the contribution of halogens to ozone loss, Read *et al.* used computer models to try to reproduce the observed concentrations of ozone[1] and its diurnal variation in the marine boundary layer around Cape Verde. Only when they included halogen compounds in their models, at the levels measured in their field study, did they succeed. If halogen compounds were disregarded, the annual average of daily ozone loss was underestimated by about 50%, and ozone concentrations were overestimated

NATURE|Vol 453|26 June 2008

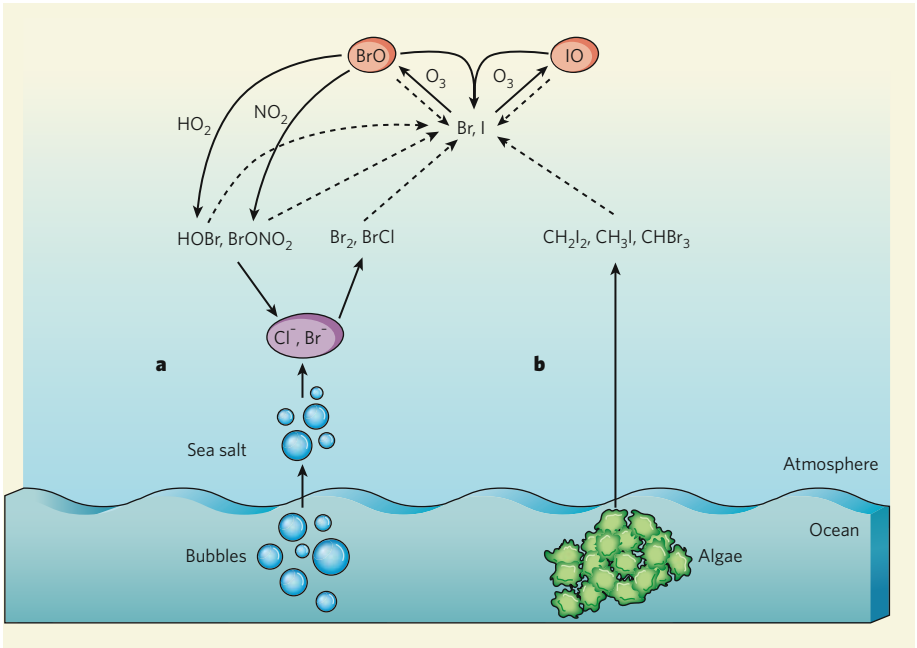


Figure 1 | Halogen release and related ozone depletion above the oceans. a, Sea water contains chloride and bromide salts that are released into the air in droplets of spray. Here they can be converted into bromine (Br_2) and bromine chloride (BrCl), which are broken up into atoms by sunlight. Bromine atoms (Br) can then react with ozone (O_3) to form bromine oxide (BrO). Dotted arrows indicate sunlight-induced reactions. **b**, Marine algae produce organic, iodine-containing molecules that are split by sunlight to produce iodine atoms (I). These then decrease ozone using the same sequence of reactions as for bromine atoms (for simplicity, not all the reactions are shown). Both BrO and iodine oxide (IO) are rapidly broken down by sunlight, re-forming ozone; for net ozone destruction, they must react with another halogen oxide, with a hydroperoxy radical (HO_2) or with nitrogen dioxide (NO_2). Read *et al.*[1] report the first long-term study of halogen oxide and ozone concentrations at a site that might be representative of Earth's open oceans.

by 12%. This is a crucial result, as most current assessments of ozone in the marine boundary layer and the troposphere do not take into account the effects of halogen chemistry, and might therefore overestimate ozone concentration, or the contributions of ozone sources.

So where do the halogen oxides come from? Ultimately, the source is the ocean. Ocean water is salty owing to the presence of sodium chloride. Bromide is also present in the ocean, but is a lot less abundant than chloride, and iodine compounds are even rarer. Droplets of sea-spray, which consist mainly of ocean water, are continuously emitted from the ocean into the atmosphere. Through a series of reactions, the bromide in the sea-spray particles can be converted to bromine compounds that are essentially insoluble in water. These compounds are therefore released from the droplets into the atmosphere, where they are rapidly broken down by sunlight, releasing bromine atoms. These atoms then react with ozone to form bromine oxide (Fig. 1a).

Iodine oxide must have a different source, because the concentrations of iodine compounds in the ocean (and so in sea-spray particles) are too small to account for the observed IO concentrations. This source is commonly thought to be marine algae. The algae produce organic, iodine-containing compounds that are broken down by sunlight to form iodine atoms, which then react with ozone to produce IO (Fig. 1b). Algae also contribute bromine to the troposphere in much the same way, but in general this is arguably a less important source than sea-salt aerosol[4].

A difference in the sources of bromine and iodine might explain why the seasonal cycles of BrO and IO at Cape Verde[1] are not the same. The details of halogen oxide production could not be tested using Read and colleagues' measurements, but several previous studies[5–8] using computer models have investigated the above-mentioned scenarios. Such models yield halogen oxide concentrations that are similar to those observed on the Cape Verde islands, lending support to these hypotheses of halogen formation.

A crucial question is whether Read and colleagues' study is truly representative of the atmosphere above the global oceans. The authors can exclude the possibility that their measurements are influenced by local halogen sources, because, if that were the case, the halogen oxide concentrations would have to be about 100 times higher than those measured in order to explain the observed ozone loss. Furthermore, the authors show that the diurnal variation of ozone is the same throughout the marine boundary layer around the islands, implying that no local halogen sources are present.

Even though the observed BrO concentrations are consistent with previous computer-model studies[5–8] that assumed sea-spray was the only source of bromine, this does not exclude the possibility that biogenic organic halogens might have a role in this region, as previous measurements[9] have shown that the organic bromine concentrations in the tropical Atlantic are high. One must therefore be cautious when extrapolating Read and colleagues' results to regions that are surrounded by a less biologically productive ocean.

To learn more about the global importance of halogen oxides over the oceans will require further long-term measurements at Cape Verde and at other island sites. Moreover, to avoid the possibility that results might be skewed by either local sources of halogens or other regional effects, ship-borne studies on the open ocean are essential. One such study has already been carried out[10], but this was in a region north of the Canary Islands, where the ocean might be comparable in biological activity to that around Cape Verde. Measurements from more geographically and biologically diverse sites are therefore essential. For this, technical obstacles must be overcome in the development of ship-borne instrumentation that can detect halogen oxides at the required sensitivity. Global oceanic surveys focusing on the halogen-catalysed destruction of ozone will also have to quantify other halogen compounds, both organic and inorganic, in addition to halogen oxides, to unravel the details of the chemical cycling.

Model studies exploring the global relevance of bromine chemistry in the troposphere[4,11] have already shown that halogen reactions could change the global ozone budget by 5–20%. Read and colleagues' measurements[1] provide initial confirmation of this, and their work will undoubtedly inspire others in their efforts to improve our quantitative understanding of ozone chemistry over the tropical and global oceans. ■

Roland von Glasow is at the School of Environmental Sciences, University of East Anglia, Norwich NR4 7TJ, UK.
e-mail: r.von-glasow@uea.ac.uk

1. Read, K. A. *et al. Nature* **453,** 1232–1235 (2008).
2. Hausmann, M. & Platt, U. *J. Geophys. Res.* **99,** 25399–25413 (1994).
3. Alicke, B., Hebestreit, K., Stutz, J. & Platt, U. *Nature* **397,** 572–573 (1999).
4. Yang, X. *et al. J. Geophys. Res.* **110,** D23311 (2005).
5. Sander, R. & Crutzen, P. J. *J. Geophys. Res.* **101,** 9121–9138 (1996).
6. Vogt, R., Crutzen, P. J. & Sander, R. *Nature* **383,** 327–330 (1996).
7. Vogt, R., Sander, R., von Glasow, R. & Crutzen, P. J. *J. Atmos. Chem.* **32,** 375–395 (1996).
8. von Glasow, R., Sander, R., Bott, A. & Crutzen, P. J. *J. Geophys. Res.* **107,** 4341 (2002).
9. Quack, B. *et al. Geophys. Res. Lett.* **31,** L23S05 (2004).
10. Leser, H., Hönninger, G. & Platt, U. *Geophys. Res. Lett.* **30,** 1537 (2003).
11. von Glasow, R., von Kuhlmann, R., Lawrence, M. G., Platt, U. & Crutzen, P. J. *Atmos. Chem. Phys.* **4,** 2481–2497 (2004).

PSYCHOLOGY

Bias at the ballot box

"Could a seemingly innocuous factor, the type of polling location where people happen to be assigned to vote, actually influence how voters cast their ballots?" Jonah Berger and colleagues asked themselves this question, and went on to answer it with two types of study (J. Berger *et al. Proc. Natl Acad. Sci. USA* doi:10.1073/pnas.0711988105; 2008).

The first was an analysis of results from a general election held in Arizona in 2000, the ballot for which included a proposition to raise state sales tax from 5.0% to 5.6%, to increase education spending. Polling stations included churches, schools, community centres and government buildings.

Berger *et al.* predicted that voting in a school would produce more support for the proposition than voting in other places. Indeed it did, but not by much compared with other documented effects on voter choice such as order on the ballot paper. Nonetheless, the effect persisted through tests for various other confounding factors (for example, the possibility of a consistently different level of voter turnout at school polling locations).

The second study was a carefully run online experiment that also involved a proposed tax increase to fund schools. The 'voting environment' was manipulated by exposing participants to typical images of schools or control images. The upshot was the same, with the school images prompting greater (and apparently unconscious) support for the initiative than, for example, an image of an office.

All in all, the authors conclude that what they call contextual priming of polling location affects how people vote. They reasonably wonder whether such factors could, for example, influence voting in a church on such matters as gay marriage and stem-cell research.

But here's a thought. In the event

K. LAMARQUE/REUTERS/CORBIS

of science spending being on the political agenda, why not offer the lab as a polling station? But maybe dim that fluorescent lighting, and persuade all those bearded fellows in white coats to take the day off — or not, as the case may be.

Tim Lincoln

MOLECULAR BIOLOGY

Power sequencing

Brenton R. Graveley

Advances in DNA-sequencing technology provide unprecedented insight into the entire collection of four genomes' transcribed sequences. They herald a new era in the study of gene regulation and genome function.

Genomes are the blueprints of life: they contain all the information necessary to build and operate their hosts. But we still have much to learn about the language of DNA to interpret the billions of Gs, As, Ts and Cs, the DNA bases that spell out life. The information-containing portions of genomes are transcribed into two RNA classes: messenger RNAs, which are translated into proteins; and non-coding RNAs, which have regulatory and mechanical roles. So studying the transcribed portion of the genome — the transcriptome — significantly aids gene identification, as well as providing insight into the inner workings of the genome and the biology of an organism. Five recent papers[1–5], including one on page 1239 of this issue by Wilhelm *et al.*[1], describe how advances in DNA-sequencing technology can be harnessed to explore transcriptomes in remarkable detail.

The concept of sequencing large numbers of randomly selected mRNAs is not new. It forms the basis of the controversial, yet revolutionary, expressed sequence tag (EST) method[6], which was originally used to identify genes in the reference copy of the human genome. In this technique, genes are quickly identified through sequencing small fragments of large numbers of mRNAs. Although EST sequencing remains useful, it is relatively slow, requires considerable resources and generally cannot identify mRNAs that are expressed at low levels.

DNA microarrays are also powerful tools for transcriptome analysis. Particularly informative are tiling arrays, which are dotted with DNA sequences derived from defined intervals (for example, every 35 base pairs) throughout the genome. Fluorescently labelled RNA is then allowed to bind to the arrays, and the transcribed portions of the genome are identified by determining which DNA sequences pair with the RNA. But tiling arrays also have several shortcomings. First, they can be used only for organisms with known genome sequences. Second, their limited sensitivity, specificity and dynamic range (the ratio of the smallest to the largest fluorescent signal) make it difficult to identify low-abundance mRNAs and to distinguish between highly similar mRNA sequences. Finally, the number of DNA probes that fit on a microarray is limited, putting constraints on the minimum feasible genomic distance between the probes, and thus on the resolution at which a genome can be analysed.

Enter the trio of next-generation sequencing technologies — systems called 454 (from 454 Life Sciences), Solexa (from Illumina) and SOLiD (from ABI) — which can generate gigabases of sequence in a single experiment[7]. They differ from traditional sequencing methods in two ways. First, rather than sequencing individual DNA clones, hundreds of thousands (the 454 system) to tens of millions (Solexa and SOLiD) of DNA molecules are sequenced in parallel. Second, the sequences obtained are much shorter (25–50 nucleotides for the Illumina and ABI technologies, and 200–400 nucleotides for the 454 system) than those generated by traditional sequencing (typically more than 800 nucleotides). Matching these shorter sequences unambiguously to the reference genome is more difficult, but this is a relatively minor trade-off compared with the massive amount of total sequence generated using these technologies. The three sequencing systems have already revolutionized the study of chromatin structure, DNA-binding proteins, DNA methylation, genome organization and small RNAs[7]. But how useful they would be for studying transcriptomes was not known.

Five teams have now used a method called mRNA-Seq (Fig. 1, overleaf) to sequence, at various levels of detail, the transcriptomes of four organisms — the fission yeast *Saccharomyces pombe*[1], the budding yeast *Saccharomyces cerevisiae*[2], the plant *Arabidopsis thaliana*[3] and the laboratory mouse[4,5]. For the sequencing step, all except one of these groups used the Solexa system[1–4], and one team[5] used the SOLiD system. In each study, between 30 and 125 million sequences — 25–39 base pairs in length — were obtained. The most inclusive of these was performed by Wilhelm *et al.*[1], who generated 122 million 39-base-pair sequences for *S. pombe*, corresponding to nearly five gigabases of sequence or 250 equivalents of this organism's genome.

But how comprehensively do these analyses cover the known genes? In the one billion bases of sequence obtained for *S. cerevisiae*, only about 91% of the known genes are detected. By contrast, sequencing five billion bases of the *S. pombe* transcriptome, Wilhelm *et al.* identify 99.3% of known genes. So although 'moderate' sequencing of the transcriptome can quickly detect most genes, identification of all genes

requires extraordinarily 'deep' sequencing.

The mRNA-Seq method can also detect previously unidentified genes. In *S. pombe*, 453 new transcripts are identified, of which 427 seem to be non-coding. Similarly, in the *S. cerevisiae* transcriptome, 204 previously undetected transcripts are identified. Although these numbers sound relatively small, they are noteworthy because the organisms investigated had already been extensively studied. Undoubtedly, mRNA-Seq will also identify unknown genes in organisms that are not typically studied in the laboratory.

Genes consist of sequences called exons, which are separated by shorter sequences known as introns. After transcription, introns are spliced out of mRNA to form mature mRNA containing only exons. One limitation posed by the spacing of DNA probes in tiling arrays is that the short introns cannot be confidently identified. mRNA-Seq, by contrast, provides unparalleled resolution because, although many sequence 'reads' do not match to introns, they cover sequences at the end of the exons on either side of the intron. These reads not only identify introns but also precisely delineate the ends of exons and introns. Data on such intron-spanning reads confirm 78% and 93% of known introns in *S. cerevisiae* and *S. pombe*, respectively. Moreover, Wilhelm and colleagues discover[1] 20 new introns in *S. pombe*.

The dynamic range, sensitivity and specificity of mRNA-Seq also make it ideal for quantitatively analysing various aspects of gene regulation, including differences in transcript abundance. For example, a comparison[3] of the transcriptome of normal *A. thaliana* with the transcriptomes of three strains of this plant that are defective in different aspects of DNA methylation — a modification that regulates gene expression — reveals scores of genes, some of them new, that are differentially expressed when DNA methylation pathways are perturbed.

The efficiency of intron removal is another aspect of gene regulation that can be monitored by comparing the number of reads that span an intron with the number that span the corresponding exon–intron junctions (Fig. 2). Wilhelm *et al.*[1] compared *S. pombe* transcriptomes from proliferating cells with those of cells undergoing different stages of meiotic cell division. They identify 314 introns from 254 genes that are spliced more efficiently during meiosis than in rapidly proliferating cells; only 12 such meiotically spliced genes were previously known. Further analysis of this data set also reveals a striking correlation between transcription levels and splicing efficiency — the higher the level of transcripts, the more efficiently they are processed to mature mRNAs.

In organisms such as the mouse, exons can be spliced together in different patterns to generate several mRNA transcripts from a single gene — a process called alternative splicing. The intron-spanning reads obtained by mRNA-Seq can also be used to identify cases of alternative splicing and to quantify changes in alternative splicing that occur in different samples[4,5]. For example, a comparison of the mRNA-Seq transcriptomes obtained from mouse brain and muscle tissues singles out[4] an exon in the *Mef2d* gene that is spliced in a specific way only in the muscle.

For transcriptome mining by mRNA-Seq, this is just the beginning of things to come. The method will become even more powerful with technological improvements such as longer reads, paired-end reads (the ability to obtain sequence from both ends of each DNA molecule and to determine the distance between those sequences)[8], enrichment for sequences of interest[9], DNA-strand-specific sequencing of the mRNA transcripts[1], and methods to sequence all RNAs[1] and not just mRNA. Algorithms that can accurately assemble short-sequence reads into longer stretches[10] will further allow sequencing of the transcriptome of organisms for which a reference genome is not available. Together, these advances will provide even greater insight into transcriptional landscapes, regulation of gene expression and alternative splicing. Most importantly, next-generation sequencing has the potential to turn individual laboratories into small genome centres and to allow an individual scientist to determine the entire transcriptome of any source (any organism, tumour samples, tissues from patients with neurodegenerative disorders, and so on) in a matter of days, and for only a few thousand US dollars. This technology will have a lasting impact on the methods and speed with which we do science. ■

Brenton R. Graveley is in the Department of Genetics and Developmental Biology, University of Connecticut Stem Cell Institute, University of Connecticut Health Center, Farmington, Connecticut 06030, USA.
e-mail: graveley@neuron.uchc.edu

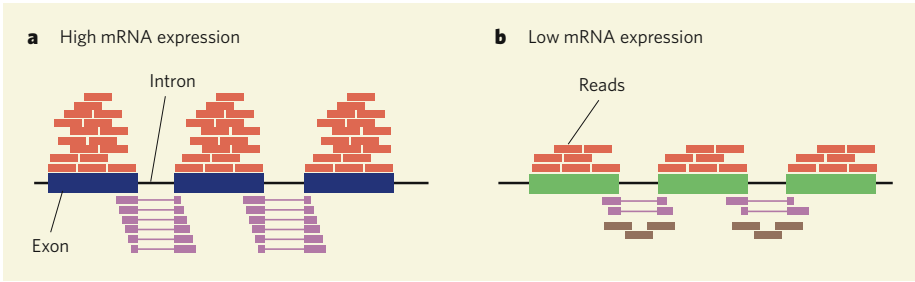


Figure 2 | Determining mRNA expression levels with mRNA-Seq. a, For a gene that is highly expressed, several sequence reads (orange) map to each of its exons and some reads (pink) map to the two exons spanning an intron. No or few reads map to introns, suggesting that intron removal in this case is efficient. **b**, But when a gene is expressed at low levels, fewer sequence reads map to the exons or to two exons spanning an intron, whereas some (brown) map to the introns. That almost equal numbers of reads map to regions within the intron and to those spanning it suggests that splicing of these gene transcripts is inefficient.

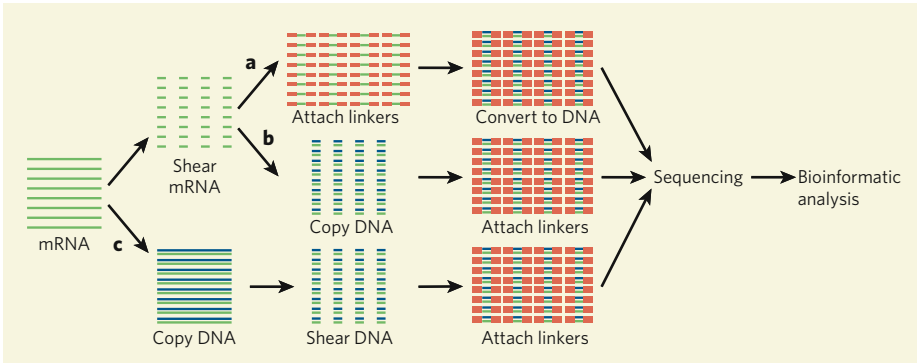


Figure 1 | mRNA-Seq. In this technique, which was used for the analysis of transcriptomes of four organisms[1–5], the isolated mRNA is analysed by one of three procedures. **a**, In the first method, mRNAs are sheared randomly, linker molecules are attached to their ends, and they are then converted to DNA. **b**, Alternatively, after shearing, the mRNA fragments are converted to DNA, and linker molecules are attached. **c**, In a third procedure, mRNAs are first copied into DNA sequences, which are then randomly sheared and attached to linkers. In all three cases, the resulting DNA is analysed by next-generation sequencing technologies and the data are compared with the reference genome for that particular organism using bioinformatics, to determine the genomic regions from which the sequences were derived.

1. Wilhelm, B. T. *et al. Nature* **453,** 1239–1243 (2008).
2. Nagalakshmi, U. *et al. Science* doi:10.1126/science.1158441 (2008).
3. Lister, R. *et al. Cell* **133,** 523–536 (2008).
4. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. *Nature Methods* doi:10.1038/nmeth.1226 (2008).
5. Cloonan, N. *et al. Nature Methods* doi:10.1038/nmeth.1223 (2008).
6. Adams, M. D. *et al. Science* **252,** 1651–1656 (1991).
7. Wold, B. & Myers, R. M. *Nature Methods* **5,** 19–21 (2008).
8. Campbell, P. J. *et al. Nature Genet.* **40,** 722–729 (2008).
9. Hodges, E. *et al. Nature Genet.* **39,** 1522–1527 (2007).
10. Zerbino, D. R. & Birney, E. *Genome Res.* **18,** 821–829 (2008).

ARTICLES

Ventastega curonica and the origin of tetrapod morphology

Per E. Ahlberg[1], Jennifer A. Clack[2], Ervīns Lukševičs[3], Henning Blom[1] & Ivars Zupiņš[4]

The gap in our understanding of the evolutionary transition from fish to tetrapod is beginning to close thanks to the discovery of new intermediate forms such as *Tiktaalik roseae*. Here we narrow it further by presenting the skull, exceptionally preserved braincase, shoulder girdle and partial pelvis of *Ventastega curonica* from the Late Devonian of Latvia, a transitional intermediate form between the 'elpistostegids' *Panderichthys* and *Tiktaalik* and the Devonian tetrapods (limbed vertebrates) *Acanthostega* and *Ichthyostega*. *Ventastega* is the most primitive Devonian tetrapod represented by extensive remains, and casts light on a part of the phylogeny otherwise only represented by fragmentary taxa: it illuminates the origin of principal tetrapod structures and the extent of morphological diversity among the transitional forms.

The fossil record of Devonian tetrapods, the earliest and most primitive limb-bearing members of the tetrapod stem group, was for many decades restricted to the iconic 'four-legged fish' *Ichthyostega* from the Famennian (latest Devonian) of Greenland[1–5] and the fragmentary genus *Acanthostega* from the same strata[2]. During the last 20 years, intense collecting and research has produced complete skeletal material of *Acanthostega*[6–8] and a series of new taxa, greatly expanding the temporal and geographical range of these animals. Devonian tetrapods are now known from as early as the late Frasnian, the earlier part of the Late Devonian period, and have been recorded from Gondwana and north China as well as Laurussia[9–18]. However, most of these new forms remain very poorly known, typically represented by no more than lower jaw rami or isolated postcranial bones; *Acanthostega* and *Ichthyostega* are still the only Devonian tetrapods known from near-complete skeletons. We know less about the fish–tetrapod transition than the taxic diversity suggests.

Among the more fragmentary forms are five (*Metaxygnathus*, *Densignathus*, *Elginerpeton*, *Obruchevichthys* and *Ventastega*) that combine a characteristically tetrapod lower-jaw morphology with the retention of coronoid fangs and other 'fish' characters absent in *Acanthostega*, *Ichthyostega* and more crownward limbed members of the tetrapod stem group[19,20]. These genera seem to fall into the morphological gap between *Acanthostega* and *Ichthyostega* and the (paraphyletic) elpistostegids, but all except *Ventastega* are very incomplete. *Ventastega* was originally described in 1994 from the Pavāri locality in the late Famennian Ketleri Formation of Kurzeme, western Latvia[21] (Supplementary Information 1). Further excavations at this site up to 2001 have yielded an extensive body of material, including previously unknown or incompletely known elements such as a near-complete skull roof plus braincase and associated cheek (Fig. 1), scapulocoracoid, anocleithrum, interclavicle and ilium (Fig. 2). All come from a single horizon, and the occurrence of multiple identical examples of several elements (jaws, cheek plates, maxillae, clavicles, cleithra, nasals) indicates that only one tetrapod taxon is present. The new material allows us to reconstruct the whole skull except the basioccipital–exoccipital complex for the first time, as well as most of the shoulder girdle and part of the pelvis (Fig. 3). It also permits a more robust phylogenetic analysis of *Ventastega*, confirming its position below *Acanthostega* in the tetrapod stem group. *Ventastega* thus provides the first detailed picture of a Devonian tetrapod more primitive than *Acanthostega*.

The skull

The overall skull shape is characteristically 'early tetrapod' with a spade-shaped snout and large dorsally positioned orbits (Figs 1 and 3a–d). However, its proportions resemble more closely those of *Tiktaalik*[22] than do the skulls of *Ichthyostega*[3] and *Acanthostega*[8], as shown both by visual comparison (Fig. 4a–c) and morphometric analysis (Fig. 4e, f and Supplementary Information 2). Furthermore, the conservation of morphological landmarks such as notches and projections of the skull-table margin is almost perfect between *Tiktaalik* and *Ventastega*, showing that the two differ only in proportions, whereas *Acanthostega* and *Ichthyostega* lack many of the landmarks. One landmark is a lateral projection posterior to the orbit, which in *Ventastega* is formed by the lateral corner of the intertemporal bone; we infer, from the presence of an identical projection in *Tiktaalik*, that an intertemporal may also be present in that genus. These results corroborate the hypothesis that the remodelling of the dermal skull across the fish–tetrapod transition was gradual[23]. The dermal skull morphology of *Tiktaalik* is closer to *Ventastega* than to the less crownward elpistostegid *Panderichthys*[24]. *Ventastega* differs from *Tiktaalik* principally in having a smaller skull table, wider spiracles and larger eyes.

As regards the dermal bone pattern of the skull (Fig. 3b–d), *Ventastega* resembles *Acanthostega* and *Ichthyostega* in retaining a preopercular bone in the cheek, but differs in possessing an intertemporal bone[3,8]. Other features are shared with *Acanthostega* but not *Ichthyostega*: these include a pair of median rostrals (also present in *Elpistostege*[23]) rather than a single bone, paired postparietals, and midline separation of the nasals. The last feature is associated in *Ventastega* with a large internasal fontanelle (Fig. 3c) which forms part of a trough-shaped midline depression in the snout. In *Acanthostega* there is only a narrow slit between the nasals and the trough is correspondingly smaller[8]. A possibly homologous small interpremaxillary fontanelle is present in several Carboniferous forms such as *Crassigyrinus*[25] and colosteids (J.A.C. personal observation) but it is unambiguously absent in *Ichthyostega*[3]. The presence of a fontanelle in *Ventastega* is clearly derived in the sense that less

[1]Subdepartment of Evolutionary Organismal Biology, Department of Physiology and Developmental Biology, Uppsala University, Norbyvägen 18A, 752 36 Uppsala, Sweden. [2]University Museum of Zoology, Cambridge, Downing Street, Cambridge CB2 3EJ, UK. [3]Department of Geology, University of Latvia, Rainis Blvd 19, Riga LV-1586, Latvia. [4]Natural History Museum of Latvia, K. Barona Str. 4, Riga LV-1712, Latvia.

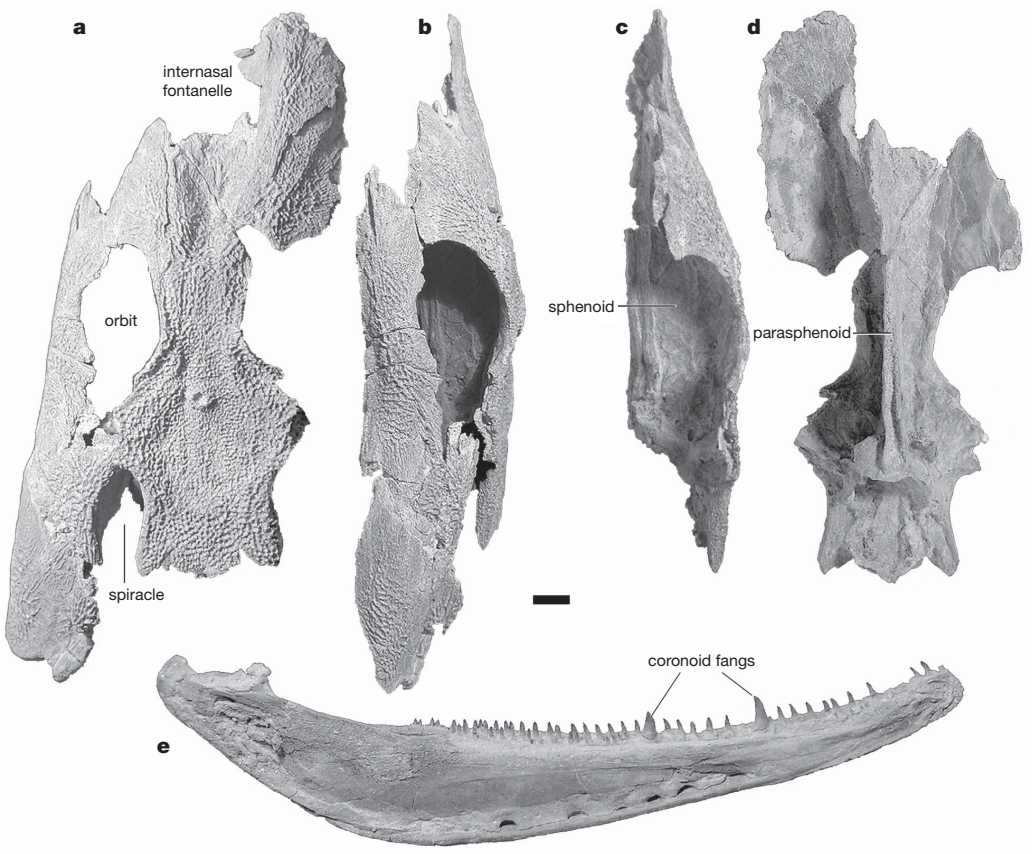


Figure 1 | Cranial material of *Ventastega*. **a**, **b**, Associated skull roof (LDM G 81/775) and cheek (LDM G 81/776) in dorsal (**a**) and left lateral (**b**) views, anterior at the top. The internasal fontanelle, orbit and spiracle are indicated in **a**. **c**, **d**, The same specimen without the cheek in left lateral (**c**) and ventral (**d**) views, anterior at the top, showing the three-dimensionally preserved braincase. The parasphenoid and sphenoid are indicated. **e**, Complete lower jaw (LDM G 81/777) in medial view with coronoid fangs shown. Scale bar, 10 mm. 'LDM G' denotes the geology collections of Latvijas Dabas Muzejs, the Natural History Museum of Latvia. For other cranial material see ref. 21.

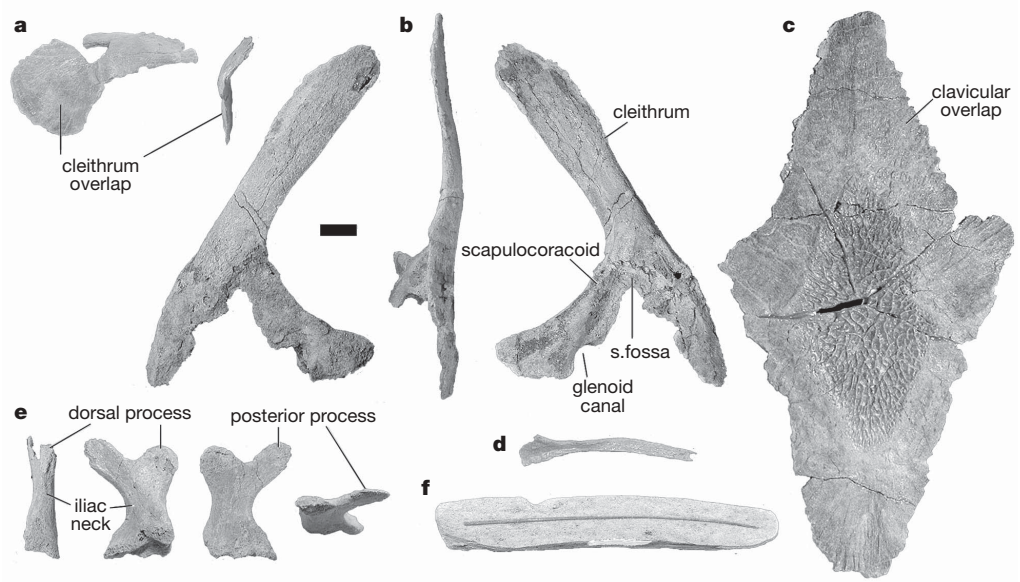


Figure 2 | Postcranial material of *Ventastega*. **a**, Right anocleithrum (LDM G 81/778) in lateral and anterior views (from left to right), showing overlap for cleithrum. **b**, Left cleithrum (LDM G 81/779) and partial scapulocoracoid in lateral, anterior and mesial views (from left to right). Note the broad shallow subscapular fossa (s.fossa) and the partially preserved glenoid canal. **c**, Interclavicle (LDM G 81/601) in ventral view showing clavicular overlaps. **d**, A probable tetrapod rib (LDM G 81/781). **e**, Right ilium (LDM G 81/780) in anterior, lateral, mesial and dorsal views (from left to right), showing the iliac neck, dorsal process and posterior process. **f**, A probable tetrapod caudal fin lepidotrichium (LDM G 81/782) on a block of matrix. Scale bar, 10 mm; all specimens shown to same scale. For other postcranial material see ref. 21.

crownward taxa like *Tiktaalik*, *Panderichthys* and tristichopterids have unbroken dermal skull roofs, but the nasal bones of these forms are separated in the midline by postrostral bone(s)[23,24]. It is thus possible that the absence of nasal–nasal contact in *Ventastega* and *Acanthostega* is primitive, with the fontanelle resulting directly from the loss of the postrostral bones. Another unique skull character of *Ventastega* is the size of the spiracular notch, which is substantially larger than those of both elpistostegids[22,26,27] and known Devonian tetrapods[3,8]. A lamina extending down from the dorsal margin of the squamosal forms part of the lateral wall of this notch. The posterior ramus of the pterygoid is narrow as in *Acanthostega*, indicating the same type of spiracular architecture[21,27]. The increase in size of the spiracular opening across the transition has been interpreted to indicate increased reliance on air-breathing among the tetrapod stem members[27–29].

The exceptionally preserved, three-dimensional braincase of *Ventastega* comprises a sphenoid and prootic region together with the dorsal part of the opisthotic (Fig. 5). The roof of the cranial cavity, spaces for the anterior and posterior semicircular canals, and endolymphatic ducts can be seen in ventral view. The basioccipital–exoccipital complex is missing, and the ethmoid region is unossified as in other early tetrapods. In most regards the braincase closely resembles that of *Acanthostega*[7]: the shape of the prootic region and its relationship to the ventral cranial fissure and the fenestra vestibuli are almost identical, as are the basipterygoid processes and the laterally open post-temporal fossae. A minor change in interpretation concerns a large and (in *Ventastega*) bi-lobed nerve foramen on the anterior face of the prootic; this was interpreted as transmitting nerve VII in *Acanthostega*[7], but its large size, position on the anterior face of the otoccipital, and bilobed shape all suggest that it is actually the opening for nerve V. The presence of a fenestra vestibuli and absence of a lateral commissure suggest that the dorsal-most element of the hyoid arch was a stapes, rather than a hyomandibula as seen in *Panderichthys*[24,27,30] and *Tiktaalik*[22]. Compared to the overall similarity between *Ventastega* and *Acanthostega*, the otoccipital region of *Ichthyostega* is very distinctive and evidently autapomorphic[4].

The one area where the braincase of *Ventastega* differs notably from that of *Acanthostega* is the orbito-temporal region immediately dorsal to the basipterygoid processes (Fig. 5b). Here, *Acanthostega* has a fairly large interorbital foramen comparable to that in many other early tetrapods[7], but *Ventastega* has a solid interorbital wall pierced only by small foramina for the pituitary vein and carotid artery, as in *Panderichthys* or 'osteolepiform' fishes—less crownward members of the tetrapod stem group[30–32]. *Ventastega* also has an anterodorsally directed tract for the optic nerve (II) with an oblique anteriorly facing opening, virtually identical to that in *Panderichthys*. *Ventastega* is more primitive than *Acanthostega* in regard to these characters; unfortunately we lack comparable information for *Ichthyostega*.

Although the braincase of *Tiktaalik* has not yet been described in detail, the published figures show a basicranial fenestra and a posteriorly

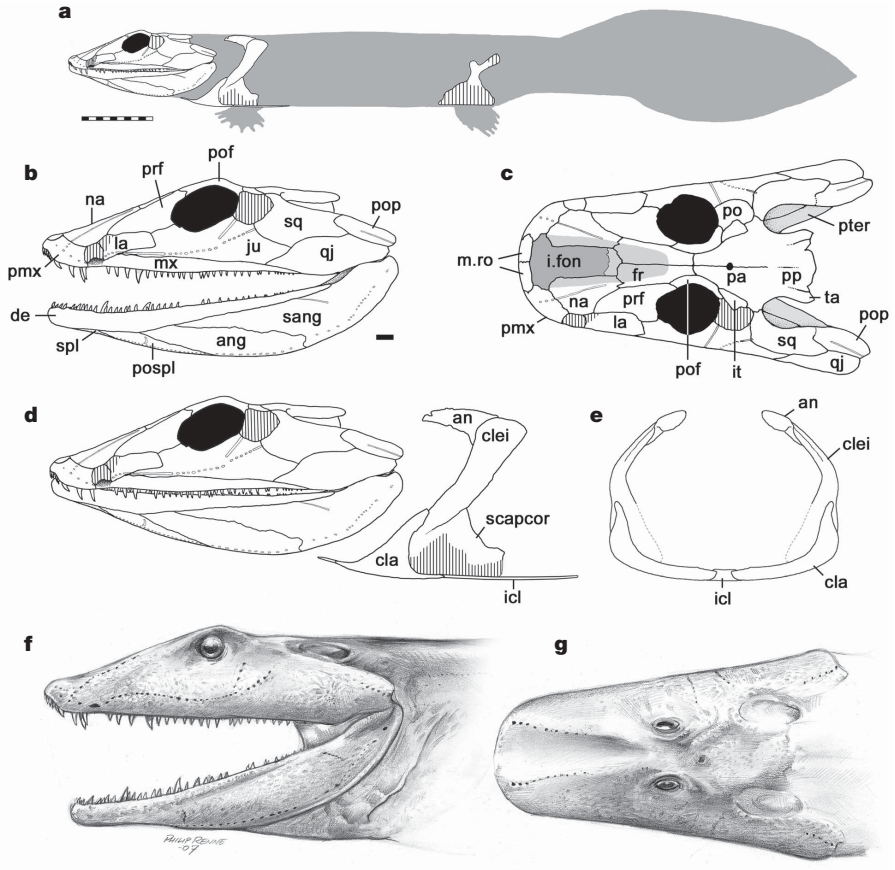


Figure 3 | Reconstructions of *Ventastega*. **a**, Whole-body reconstruction showing known skeletal elements on a body outline based on *Acanthostega* (modified from ref. 5; original *Acanthostega* body reconstruction by M. I. Coates). Scale bar, 10 cm. **b**, **c**, Skull reconstruction in lateral and dorsal views, based on material presented here and described previously[21]. **d**, Reconstructed association of skull and shoulder girdle in lateral view. **e**, Shoulder girdle in anterior view. Curvature of cleithrum based on LDM G 81/522 (ref. 21). Unknown bones are indicated with vertical hatching. Scale bar for **b**–**e**, 10 mm. **f**, **g**, Life reconstructions of head in lateral and dorsal views (copyright P. Renne, 2007). an, anocleithrum; ang, angular; cla, clavicle; clei, cleithrum; de, dentary; fr, frontal; icl, interclavicle; i.fon, internasal fontanelle; it, intertemporal; ju, jugal; la, lacrimal; mx, maxilla; m.ro, median rostral; na, nasal; pa, parietal; pmx, premaxilla; po, postorbital; pof, postfrontal; pop, preopercular; pospl, postsplenial; pp, postparietal; prf, prefrontal; pter, pterygoid; qj, quadratojugal; sang, surangular; scapcor, scapulocoracoid; spl, splenial; sq, squamosal; ta, tabular.

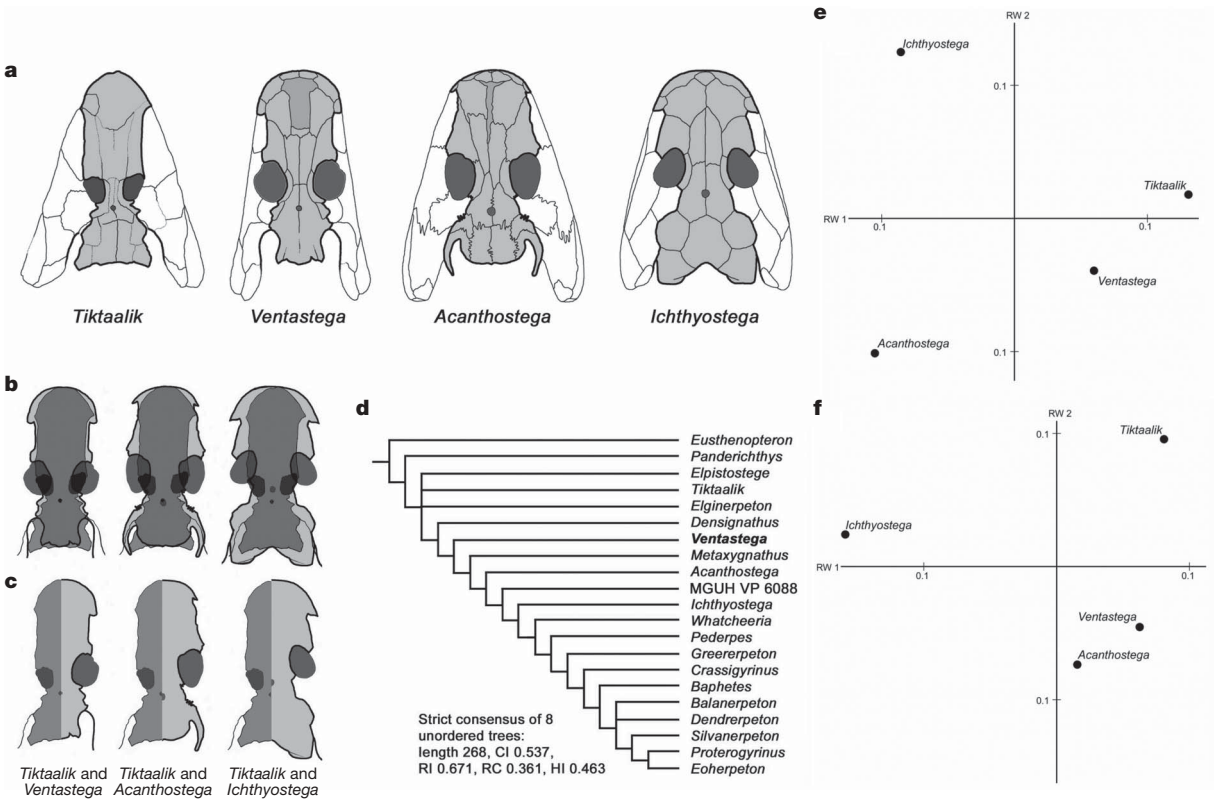


Figure 4 | Skull shape and phylogeny. **a**, Skulls of *Tiktaalik*, *Ventastega*, *Acanthostega* and *Ichthyostega* in dorsal view, showing the skull roof (grey) used in the morphometric comparison. In *Ventastega* and *Acanthostega* the internasal fontanelle is shown darker grey. Not drawn to scale. **b**, **c**, Comparison of the skull roofs of *Tiktaalik* and *Ventastega* (left), *Tiktaalik* and *Acanthostega* (centre) and *Tiktaalik* and *Ichthyostega* (right). The skull roofs are overlaid in **b**; a left half-roof of *Tiktaalik* is compared to a right half-roof of *Ventastega*, *Acanthostega* or *Ichthyostega* in **c**. *Tiktaalik* is shown in darker grey than the tetrapods. A slight distortion of *Tiktaalik* has been corrected using the 'skew' command in Photoshop (**b**, **c**). **d**, Strict consensus unordered phylogeny of tetrapodomorph fishes and early tetrapods based on 117 characters scored for 21 taxa. For further phylogenies see Supplementary Information 3. 'MGUH VP 6088' is an undescribed Famennian tetrapod from Greenland. CI, consistency index; HI, homoplasy index; RC, rescaled consistency index; RI, retention index. **e**, **f**, Relative warp analyses of skull roof outlines shown in **a**–**c**; including (**e**) and excluding (**f**) the tabular horn of *Acanthostega*. The first relative warp (RW) is on the horizontal axis; the second relative warp is on the vertical axis. For a full discussion of the relative warp analysis see Supplementary Information 2.

positioned lateral commissure supporting a hyomandibula[22]. These features compare closely with *Panderichthys*[30], probably indicating a broadly similar morphology—a 'lobe-fin' otoccipital comparable at least in its ventral parts to *Eusthenopteron*[31] or *Gogonasus*[32] but different from the tetrapod pattern. *Tiktaalik* also retains pterygoid separation by the parasphenoid and an osteolepiform lower jaw structure, whereas

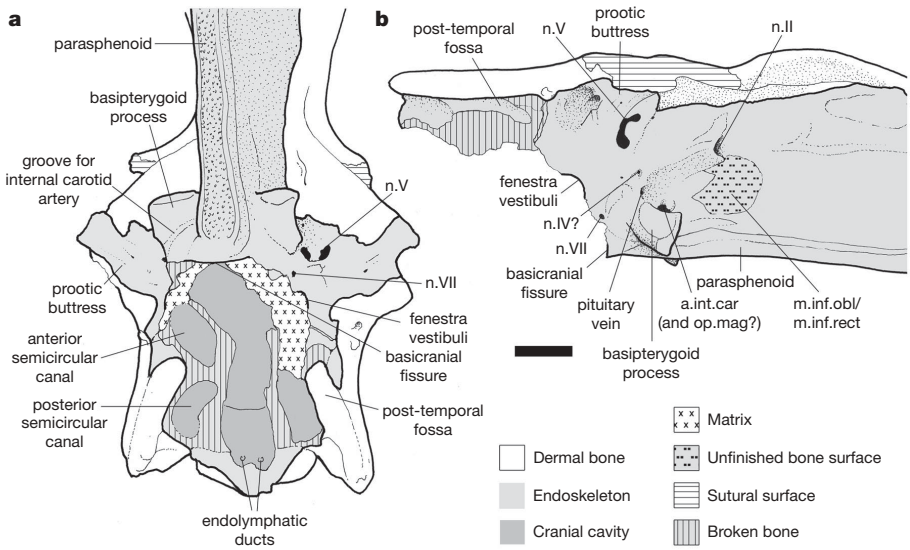


Figure 5 | Braincase of *Ventastega*. **a**, **b**, Posterior half of braincase plus skull roof of LDM G 81/775 in ventral and lateral views. a.int.car, foramen for the internal carotid artery; m.inf.obl/m.inf.rect, muscle scar for the inferior obliquus and/or inferior rectus eye muscles; n.II–n.VII, cranial nerves II, IV, V and VII; op.mag, ophthalmica magna artery. Scale bar, 10 mm.

Ventastega has pterygoid–pterygoid contact and a tetrapod lower jaw albeit with coronoid fangs[19,21,22].

The postcranial skeleton

The preserved pectoral girdle of *Ventastega* comprises interclavicle, clavicle, cleithrum, scapulocoracoid and anocleithrum (Figs 2 and 3d, e); the cleithrum was previously misidentified as an ilium[21], but a real ilium has since been collected and the discovery of a cleithrum with attached scapulocoracoid confirms its identity. Of these elements the interclavicle, clavicle and anocleithrum resemble those of *Acanthostega*[6] rather than *Ichthyostega*. The cleithrum of *Ventastega* is similar in outline to those of *Ichthyostega* and *Acanthostega*, and like them it lacks ornament, but in contrast to *Acanthostega* it lacks a postbranchial lamina. Such a lamina is also absent in *Tiktaalik*[33], suggesting that its presence in *Acanthostega* may not be primitive as originally supposed[34]. The scapulocoracoid of *Ventastega* is incomplete (Fig. 2b), but enough is preserved to show that it is essentially *Acanthostega*-like with a broad shallow subscapular fossa[6]. In *Ichthyostega*[3], *Hynerpeton*[14] and the girdles attributed to *Elginerpeton*[13], the subscapular fossa is deeper with a more acute apex. A large, posteriorly positioned, partly preserved foramen in the scapulocoracoid of *Ventastega* may correspond to the 'glenoid canal' of *Ichthyostega*[3] and foramina 'D' and 'E' (or possibly 'A') of *Acanthostega*[6]. There is no trace of a coracoid foramen similar to that in *Tiktaalik*[33]. As in all Devonian tetrapods except *Tulerpeton*[11], a scapular blade is absent. Overall, the pectoral girdle of *Ventastega* is clearly of tetrapod grade, quite different from those of *Panderichthys*[35] and *Tiktaalik*, and we infer that it bore limbs with digits.

An incomplete right ilium of *Ventastega* (Fig. 2e) also shows an *Acanthostega*-like morphology[6]: the slender iliac neck—which lacks an iliac canal—branches into a distinct dorsal process with an unfinished dorsal surface and a posterodorsally directed posterior process with an upright oval cross-section. In *Ichthyostega*, by contrast, the robust iliac neck is pierced by a canal, the dorsal process is broader and less distinct, and the posterior process is horizontal[3]. These characters also occur in the ilia attributed to *Elginerpeton*[13]. In addition to these unambiguous stem tetrapod bones, Pavāri also yields numerous slender unjointed lepidotrichia, 70 mm or more in length (Fig. 2f), which we tentatively interpret as caudal lepidotrichia of *Ventastega* because of their similarity to those of *Acanthostega*[6]. A single slender *Acanthostega*-like rib (Fig. 2d) may also belong to *Ventastega*. The strongly *Acanthostega*-like character of the postcranial bones, coupled with the evidence for a large caudal fin, suggest that the overall body morphology of *Ventastega* resembled *Acanthostega*. We have accordingly used a reconstructed body outline of *Acanthostega*[5], originally based on the work of M. I. Coates, as the basis for a tentative reconstruction of *Ventastega* (Fig. 3a).

Ventastega and the origin of tetrapods

Although *Ventastega* is one of the youngest Devonian tetrapods, deriving from the late Famennian, it occupies a relatively deep position in the tetrapod stem group. All permutations of our phylogenetic analysis (Fig. 4d and Supplementary Information 3) place it below both *Ichthyostega* and *Acanthostega*; only *Elginerpeton* consistently occupies a more basal position. The postcranial elements attributed to *Elginerpeton* show that vertebrates with limbs had originated before the end of the Frasnian[13]. The recent redating of *Metaxygnathus* as late Frasnian[36], in conjunction with the phylogenetic topologies recovered by our analysis, implies not only that *Ventastega* represents a lineage of Frasnian origin but that a substantial part of the Devonian tetrapod radiation occurred during the Frasnian. This is consistent with the occurrence of *Livoniana*, a fragmentary taxon apparently more derived than *Tiktaalik*, in the latest Givetian of the Baltic region[37]. It seems that the Famennian tetrapod record has only a poor stratophylogenetic fit, a contention that is further supported by the co-occurrence of the very primitive humerus ANSP 21350 (ref. 38) and much more derived whatcheeriid-like skull elements (J.A.C. personal observation) in the upper Famennian Catskill Formation of Pennsylvania.

Overall, the character combination shown by *Ventastega* carries a clear signal: with the exception of some possible autapomorphies, all its character states match either *Acanthostega* or the elpistostegids *Elpistostege*, *Tiktaalik* and *Panderichthys*. No characters are shared uniquely with *Ichthyostega* or with the cranial and attributed postcranial material of *Elginerpeton*. Among the less complete tetrapod stem-group members, *Metaxygnathus* and *Densignathus* have lower jaws rather similar to *Ventastega*, but their general morphology is unknown[15,19,20]. This pattern suggests that the shared *Ventastega*–*Acanthostega* character complex is paraphyletically distributed through a segment of the tetrapod stem group rather than being synapomorphies of a clade. Consistent with this interpretation is the fact that certain aspects of the character complex, for example, the shape of the otic capsule and ilium, also occur in much later and more derived tetrapods such as anthracosaurs[39,40] and *Crassigyrinus*[41]. We interpret these as persistent primitive traits rather than homoplastic reversals in the latter taxa. The morphometric similarities between *Ventastega* and *Tiktaalik*, in particular the conservation of landmarks around the skull table, suggest that the changes in skull shape during this part of the fish–tetrapod transition were substantially proportional: the eyes and spiracles grew larger, the skull table smaller, and the snout broader. This contrasts with marked pattern changes in the dermal bones of the cheek, skull roof and palate, and with a restructuring of braincase that resulted in the loss of the intracranial joint, basicranial fenestra and lateral commissure as well as a host of other smaller changes. With a few modifications such as the gradual withdrawal of the notochord and the rearward extension of the parasphenoid across the basicranial fissure, this new braincase morphology remained essentially constant up into the base of the tetrapod crown group[42]. Even the highly specialized braincase of *Ichthyostega* is recognizably derived from this pattern[4]. With regard to the postcranial skeleton, *Ventastega* consistently resembles *Acanthostega*; all the changes that distinguish Devonian tetrapod from elpistostegid limb girdles—loss of the supracleithrum and post-temporal; enlargement of the scapulocoracoid; loss of the coracoid foramen; enlargement of the interclavicle, creation of a sacrum—seem to have already occurred.

Because of its phylogenetic position and character complement it is tempting to interpret *Ventastega* as a straightforward evolutionary intermediate, which represents with reasonable accuracy the character complement of the tetrapod stem lineage at a point on the internode between *Tiktaalik* and *Acanthostega*. However, this simple picture should be approached with a degree of caution. ANSP 21350 and *Elginerpeton* in particular (whether or not the latter taxon is taken to include the disputed humerus GSM 104536; refs 13, 38) show character combinations that are substantively different from those of *Ventastega* and *Acanthostega* without being obviously autapomorphic, and both probably occupy deep positions in the phylogeny. At a minimum this demonstrates the presence of considerable morphological diversification among the earliest tetrapods. More importantly, however, the discovery of articulated material of these or similar forms could have a substantial impact on the tree topology. *Ventastega*, like *Tiktaalik*, conforms remarkably well to prior expectations of what a transitional form at that particular point in the phylogeny should be like; whether the same will be true of future discoveries remains to be seen.

METHODS SUMMARY

The material was excavated from Pavāri locality in 1970, 1973, 1988, 1991, 1995 and 2001 and deposited at the Natural History Museum of Latvia. In the laboratory, fossils were freed from surrounding sediment (unconsolidated sand) by mechanical preparation with a mounted needle. Relative warps analysis[43] was used to quantify head-shape variation in the various Devonian tetrapods and

elpistostegids. Landmarks were digitized from published reconstructions[3,8,22,24] using the program tpsDig v. 1.40 (ref. 44). Relative warps analysis was conducted in tpsRelw v. 1.39 (ref. 45). Phylogenetic analysis was performed in PAUP 4.0b10 (ref. 46) using a Branch-and-Bound search with default settings, with *Eusthenopteron* specified as the out-group. Life reconstructions were drawn by P. Renne under the supervision of P.E.A.

Received 22 November 2007; accepted 9 April 2008.

1. Säve-Söderbergh, G. Preliminary note on Devonian stegocephalians from East Greenland. *Meddr. Grønland* **94,** 1–107 (1932).
2. Jarvik, E. On the fish-like tail in the ichthyostegid stegocephalians with descriptions of a new stegocephalian and a new crossopterygian from the Upper Devonian of East Greenland. *Meddr. Grønland* **114,** 1–90 (1952).
3. Jarvik, E. The Devonian tetrapod *Ichthyostega. Fossils Strata* **40,** 1–213 (1996).
4. Clack, J. A. *et al.* A uniquely specialized ear in a very early tetrapod. *Nature* **425,** 65–69 (2003).
5. Ahlberg, P. E., Clack, J. A. & Blom, H. The axial skeleton of the Devonian tetrapod *Ichthyostega. Nature* **437,** 137–140 (2005).
6. Coates, M. I. The Devonian tetrapod *Acanthostega gunnari* Jarvik: postcranial anatomy, basal tetrapod interrelationships and patterns of skeletal evolution. *Trans. R. Soc. Edinb. Earth Sci.* **87,** 363–421 (1996).
7. Clack, J. A. The neurocranium of *Acanthostega gunnari* Jarvik and the evolution of the otic region in tetrapods. *Zool. J. Linn. Soc.* **122,** 61–97 (1998).
8. Clack, J. A. A revised reconstruction of the dermal skull roof of *Acanthostega*, an early tetrapod from the Late Devonian. *Trans. R. Soc. Edinb. Earth Sci.* **93,** 163–165 (2003).
9. Campbell, K. S. W. & Bell, M. W. A primitive amphibian from the Late Devonian of New South Wales. *Alcheringa* **1,** 369–381 (1977).
10. Lebedev, O. A. & Clack, J. A. Upper Devonian tetrapods from Andreyevka, Tula Region, Russia. *Palaeontology* **36,** 721–734 (1993).
11. Lebedev, O. A. & Coates, M. I. The postcranial skeleton of the Devonian tetrapod *Tulerpeton curtum. Zool. J. Linn. Soc.* **114,** 307–348 (1995).
12. Ahlberg, P. E. *Elginerpeton pancheni* and the earliest tetrapod clade. *Nature* **373,** 420–425 (1995).
13. Ahlberg, P. E. Postcranial stem tetrapod remains from the Devonian of Scat Craig, Morayshire, Scotland. *Zool. J. Linn. Soc.* **122,** 99–141 (1998).
14. Daeschler, E. B., Shubin, N. H., Thomson, K. S. & Amaral, W. W. A Devonian tetrapod from North America. *Science* **265,** 639–642 (1994).
15. Daeschler, E. B. Early tetrapod jaws from the Late Devonian of Pennsylvania, USA. *J. Paleontol.* **74,** 301–308 (2000).
16. Zhu, M., Ahlberg, P. E., Zhao, W. & Jia, L. First Devonian tetrapod from Asia. *Nature* **420,** 760–761 (2002).
17. Clement, G. *et al.* Devonian tetrapod from Western Europe. *Nature* **427,** 412–413 (1994).
18. Lebedev, O. A. A new tetrapod *Jakubsonia livnensis* from the Early Famennian (Devonian) of Russia and palaeoecological remarks on the Late Devonian tetrapod habitats. *Acta Univ. Latviensis* **679,** 79–98 (2004).
19. Ahlberg, P. E. & Clack, J. A. Lower jaws, lower tetrapods - a review based on the Devonian genus *Acanthostega. Trans. R. Soc. Edinb. Earth Sci.* **89,** 11–46 (1998).
20. Ahlberg, P. E., Friedman, M. & Blom, H. New light on the earliest known tetrapod jaw. *J. Vert. Paleontol.* **25,** 720–724 (2005).
21. Ahlberg, P. E., Lukševičs, E. & Lebedev, O. The first tetrapod finds from the Devonian (Upper Famennian) of Latvia. *Phil. Trans. R. Soc. B* **343,** 303–328 (1994).
22. Daeschler, E. B., Shubin, N. H. & Jenkins, F. A. A Devonian tetrapod-like fish and the evolution of the tetrapod body plan. *Nature* **440,** 757–763 (2006).
23. Schultze, H. P. & Arsenault, M. The panderichthyid fish *Elpistostege*: a close relative of tetrapods? *Palaeontology* **28,** 293–309 (1985).
24. Vorobyeva, E. I. Morphology and nature of evolution of crossopterygian fishes [In Russian]. *Trudy Paleont. Inst.* **163,** 1–239 (1977).
25. Clack, J. A. The Scottish Carboniferous tetrapod *Crassigyrinus scoticus* (Lydekker) – cranial anatomy and relationships. *Trans. R. Soc. Edinb. Earth Sci.* **88,** 127–142 (1998).
26. Vorobyeva, E. I. & Schultze, H. P. Description and systematics of panderichthyid fishes with comments on their relationship to tetrapods. In *Origins of the Higher Groups of Tetrapods* (eds Schultze, H. P. & Trueb, L.) 68–109 (Cornell, Ithaca, New York, 1991).
27. Brazeau, M. D. & Ahlberg, P. E. Tetrapod-like middle ear architecture in a Devonian fish. *Nature* **439,** 318–321 (2006).
28. Long, J. A., Young, G. C., Holland, T., Senden, T. J. & Fitzgerald, E. M. G. An exceptionally preserved Devonian fish from Australia sheds light on tetrapod origins. *Nature* **444,** 199–202 (2006).
29. Clack, J. A. Devonian climate change, breathing, and the origin of the tetrapod stem group. *Integr. Comp. Biol.* **47,** 510–523 (2007).
30. Ahlberg, P. E., Clack, J. A. & Lukševičs, E. Rapid braincase evolution between *Panderichthys* and the earliest tetrapods. *Nature* **381,** 61–64 (1996).
31. Jarvik, E. *Basic Structure and Evolution of Vertebrates* Vol. 1 (Academic, London, 1980).
32. Long, J. A., Barwick, R. E. & Campbell, K. S. W. Osteology and functional morphology of the osteolepiform fish *Gogonasus andrewsae* Long 1985, from the Upper Devonian Gogo Formation, Western Australia. *Rec. West. Aust. Mus.* **57** (Suppl.), 1–89 (1997).
33. Shubin, N. H., Daeschler, E. B. & Jenkins, F. A. The pectoral fin of *Tiktaalik rosae* and the origin of the tetrapod limb. *Nature* **440,** 764–771 (2006).
34. Coates, M. I. & Clack, J. A. Fish-like gills and breathing in the earliest known tetrapods. *Nature* **352,** 234–236 (1991).
35. Vorobyeva, E. I. The shoulder girdle of *Panderichthys rhombolepis* (Gross) (Crossopterygii), Upper Devonian, Latvia. *GeoBios* **28,** 285–288 (1995).
36. Young, G. C. Biostratigraphic and biogeographic context for tetrapod origins during the Devonian: Australian evidence. *Alcheringa* **1** (Special Issue), 409–428 (2006).
37. Ahlberg, P. E., Lukševičs, E. & Mark-Kurik, E. A near-tetrapod from the Baltic Middle Devonian. *Palaeontology* **43,** 533–548 (2000).
38. Shubin, N. H., Daeschler, E. B. & Coates, M. I. The early evolution of the tetrapod humerus. *Science* **304,** 90–93 (2004).
39. Clack, J. A. *Pholiderpeton scutigerum* Huxley, an amphibian from the Yorkshire coal measures. *Phil. Trans. R. Soc. B* **318,** 1–107 (1987).
40. Clack, J. A. & Holmes, R. The braincase of the anthracosaur *Archeria crassidisca*, with comments on the interrelationships of primitive tetrapods. *Palaeontology* **31,** 85–107 (1988).
41. Panchen, A. L. & Smithson, T. R. The pelvic girdle and hind limb of *Crassigyrinus scoticus* (Lydekker) from the Scottish Carboniferous and the origin of the tetrapod pelvic skeleton. *Trans. R. Soc. Edinb. Earth Sci.* **81,** 31–44 (1990).
42. Ruta, M., Coates, M. I. & Quicke, D. L. J. Early tetrapod relationships revisited. *Biol. Rev.* **78,** 251–345 (2003).
43. Bookstein, F. L. *Morphometric tools for landmark data: geometry and biology* (Cambridge Univ. Press, 1991).
44. Rohlf, F. J. *tpsDig, version 1.40* (Dept. of Ecology & Evolution, Stony Brook, New York, 2004).
45. Rohlf, F. J. *tpsRelw, version 1.39* (Dept. of Ecology & Evolution, Stony Brook, New York, 2004).
46. Swofford, D. *PAUP*: Phylogenetic Analysis Using Parsimony (and other methods) 4.0 Beta* (Sinauer Associates, Sunderland, Massachusetts, 2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge the hard work and dedication of the Pavāri field crews of 1988, 1991, 1995 and 2001. P.E.A. acknowledges the financial support of the Natural History Museum, London (1995 and 2001 field seasons) and the Swedish Research Council. H.B. acknowledges the support of NERC (2001 field season) and the Swedish Research Council. E.L. acknowledges the financial support of the Latvian Council of Science. Special thanks to Philip Renne for his life reconstructions.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.E.A. (per.ahlberg@ebc.uu.se).

ARTICLES

Scaling of the BMP activation gradient in *Xenopus* embryos

Danny Ben-Zvi[1], Ben-Zion Shilo[1], Abraham Fainsod[2] & Naama Barkai[1,3]

In groundbreaking experiments, Hans Spemann demonstrated that the dorsal part of the amphibian embryo can generate a well-proportioned tadpole, and that a small group of dorsal cells, the 'organizer', can induce a complete and well-proportioned twinned axis when transplanted into a host embryo. Key to organizer function is the localized secretion of inhibitors of bone morphogenetic protein (BMP), which defines a graded BMP activation profile. Although the central proteins involved in shaping this gradient are well characterized, their integrated function, and in particular how pattern scales with size, is not understood. Here we present evidence that in *Xenopus*, the BMP activity gradient is defined by a 'shuttling-based' mechanism, whereby the BMP ligands are translocated ventrally through their association with the BMP inhibitor Chordin. This shuttling, with feedback repression of the BMP ligand Admp, offers a quantitative explanation to Spemann's observations, and accounts naturally for the scaling of embryo pattern with its size.

Multicellular organisms develop with a remarkable consistency, maintaining a precise body plan in the face of genetic polymorphism or environmental fluctuations[1]. Yet, size and shape differ significantly even between closely related species. Developmental processes are thus shaped by seemingly opposing challenges: maintaining robustness at the species level, while allowing sufficient flexibility for evolutionary adaptation[2]. The interplay between robustness and evolutionary plasticity is poorly understood.

In Bilateria, early dorsoventral patterning relies on the graded distribution of BMP activity along the embryo. Two classical experiments performed by Hans Spemann (reviewed in ref. 3) demonstrated the dramatic plasticity of this patterning process in amphibians. First, dorsal halves of bisected embryos develop into well-proportioned tadpoles[4]. Second, cells taken from the embryonic dorsal blastopore lip and transplanted into the ventral side of a naive embryo induce a complete and well-proportioned secondary axis[5]. The region responsible for this induction property, 'Spemann's organizer', was identified later in other vertebrates, and its inductive capacity is attributed primarily to the secretion of BMP inhibitors[6]. However, the mechanism underlying the ability of dorsal-half embryos to grow into well-scaled tadpoles, and the ability to generate two complete and properly scaled tadpoles upon organizer transplantation, remained unknown. Experiments by Cooke further demonstrated the precision of scaling, and verified that compensation is not due to overgrowth of the remaining cells, but to their proportionate assignment to the different tissues[7].

Despite large differences in shape and size, the molecular network that generates the BMP gradient is remarkably conserved across evolution[8–11]. In flies and vertebrates, BMP ligands are initially expressed in broad domains, with the localized secretion of a conserved BMP inhibitor (Sog/Chordin, respectively[12,13]) providing the key organizing dorsoventral asymmetry[14,15] (Supplementary Fig. 1a–d). The inhibitors diffuse, undergo cleavage by a conserved protease (Tld/Xlr) and interact with a conserved modulator (Tsg/xTsg)[16–21]. Anti-dorsalizing morphogenic protein (Admp) is a BMP ligand found in many Bilateria but is missing in *Drosophila*. In contrast to other BMP ligands, it is expressed dorsally with BMP inhibitors, and is subject to autoregulatory transcriptional repression by the BMP pathway[22,23]. A role for Admp in providing scaling was recently suggested, following the observation that depletion of Admp abolishes patterning in dorsal-half embryos[24].

Theoretical analysis of the BMP gradient formation in the *Drosophila* embryo distinguished two qualitatively different patterning mechanisms[25–28] (Supplementary Fig. 1e, f). In the 'inhibition-based' mechanism, patterning is governed by the creation of an inhibition gradient over a uniform field of activators. In the 'shuttling-based' mechanism, patterning relies on the physical translocation of the activator to the midline, mediated by its binding to the inhibitor. Both mechanisms can generate a graded profile of BMP activation[26,29], but the finding that the shuttling mechanism generates a sharp and robust gradient led to the proposal that it is in use. This prediction was subsequently verified experimentally[26,30–32] (reviewed in refs 33 and 34). In this study, we show that shuttling is used also in the *Xenopus* embryo, and that, with the auto-repression of Admp, it ensures the scaling of the BMP activation profile with embryo size.

Shuttling is required for scaling

A key question is whether the conservation of network constituents implies the conservation of their integrated function, and if so, how an increased functional complexity can evolve. To examine whether shuttling plays a part in establishing the BMP activation gradient in *Xenopus*, we focused first on the ability of dorsal-half embryos to generate a well-proportioned embryo. This scaling property is difficult to explain by most models of morphogen gradients, and was proposed as evidence that patterning does not involve morphogens[7]. To assess the constraints imposed by this scaling property rigorously, we formulated a mathematical model that is based on the conserved core of this patterning network. The model includes two BMP ligands, Admp and Bmp (where Bmp stands for the three ligands BMP2/4/7), a BMP inhibitor (Chordin) and the protease Xlr (Fig. 1a). We allowed for the diffusion of all components, the binding of the BMP ligands to Chordin, and the degradation of Chordin by Xlr. We also considered the production of Chordin and Admp at the dorsal pole and the auto-repression of Admp by BMP signalling[24] (see Methods and Supplementary Information for equations and further details of the screen).

[1]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel. [2]Department of Cellular Biochemistry and Human Genetics, Faculty of Medicine, Hebrew University, Jerusalem 91120, Israel. [3]Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel.

We screened systematically for parameters (rate constants and diffusion coefficients) for which the activation gradient is robust and scales with embryo size. To this end, we assigned each of the nine parameters three possible values, ranging over at least two orders of magnitude. The BMP activation gradient was solved numerically for all the networks defined in this nine-dimensional cube. Of the 26,000 networks examined, approximately 1,100 displayed a proper polarity, but only 21 were also capable of scaling in a dorsal-half embryo. An example for a gradient that did not scale is shown in Fig. 1b, d whereas a scaled gradient is shown in Fig. 1c, e. Examining the solutions, we noted that the respective gradients were established by distinct mechanisms. In the network that did not scale, the overall level of the ligands remained approximately uniform, and the activation gradient reflected the gradient of the inhibitor Chordin. In contrast, in the network that did scale, the ligands were physically concentrated at the ventral pole. The activation gradient was thus generated by the shuttling of the ligands to the ventral pole.

To examine more generally whether shuttling is required for scaling, we defined a rigorous measure that quantified the extent of shuttling, namely the translocation of the total ligand to the ventral pole, in each of the networks tested (Methods and Supplementary Information). Shuttling was observed in all networks that were capable of robust scaling (Fig. 1f). Moreover, the parameters of the consistent networks obeyed the molecular requirement for shuttling, as described previously in *Drosophila*[26]. First, the binding of Chordin to the BMP ligands largely facilitated the diffusion of the BMP ligands (Fig. 1g–j). Second, Chordin was degraded primarily when complexed with BMP ligands (Fig. 1g, h). Additionally, we found that scaling requires that Chordin binds to Bmp with a significantly higher affinity than to Admp (Fig. 1i, j).

Mechanism by which scaling is achieved

Our numerical analysis confirmed that the known patterning network can support the scaling of pattern with size, and suggested that shuttling plays an important part in providing this ability. To

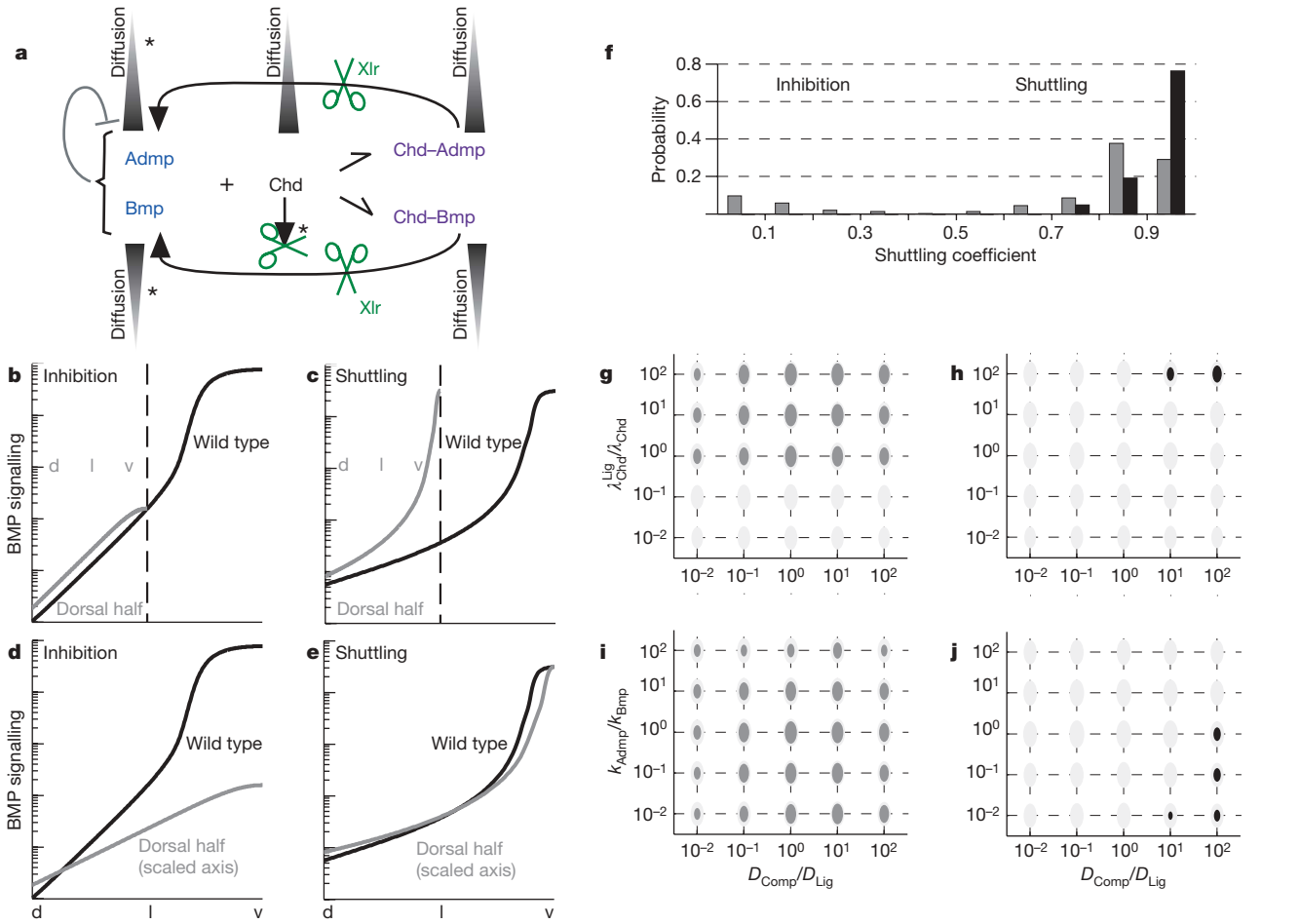


Figure 1 | Numerical evidence for shuttling. a, Model used in the screen. See Methods and Supplementary Information for equations and parameters. **b**, **d**, Activation profiles defined by the inhibition-based model. The model assumes that all components diffuse at the same rate. The model was solved numerically for the whole embryos (black) and the dorsal-half embryos (grey). The unscaled profile, measured in absolute length, is shown in **b** whereas the scaled profile measured in relative lengths, scaled by the size of the field, is shown in **d**. d, l, v, dorsal, lateral and ventral regions of the embryo, respectively. **c**, **e**, Activation profiles defined by the shuttling-based model. The model assumes free-ligand diffusion is much smaller than that of Chordin and the complex. Chordin is degraded primarily when complexed with a BMP ligand (parameters marked with * in Fig. 1a are small). The unscaled profile is shown in **c** whereas the scaled profile is shown in **e**. **f**, Consistent networks establish pattern by shuttling. Distribution of the shuttling coefficient S_h in consistent networks (black) and in all the networks that establish a dorsoventral gradient (grey). $S_h \approx 1$ indicates a shuttling mechanism, whereas $S_h \approx 0$ an inhibition mechanism. **g**, **h**, Relative diffusion and degradation in networks that establish dorsoventral polarity. The *x* axis displays the ratio between diffusion coefficients of the inhibitor-bound and free BMP ligand (D_{Comp}/D_{Lig}). The *y* axis displays the ratio between degradation of BMP-bound and free inhibitor $\left(\lambda^{Lig}_{Chd}/\lambda_{Chd}\right)$. Networks that were sampled in the screen are in light grey. Because this plot is a projection from a nine-dimensional space, each point symbolizes many networks where the respective ratios were held fixed, but parameters were changed systematically. **g**, Grey circles correspond to networks establishing proper dorsoventral polarity. **h**, Black circles correspond to networks establishing proper dorsoventral polarity, support scaling and robustness. **i**, **j**, Affinity of Chordin to Admp versus Bmp. Same as **g**–**h** with the *y* axis denoting the ratio between of binding rates of Chordin to Admp and Bmp (k_{Admp}/k_{Bmp}).

understand better how scaling is achieved we analysed the shuttling mechanism in more detail (Supplementary Information). As we have shown previously[26], shuttling of a single ligand leads to an activation profile that decays as a power law:

$$S(x) \approx \frac{S_0^{\mathrm{Lig}}}{x^2}; \quad S_0^{\mathrm{Lig}} = \frac{2D_{\mathrm{Chd}}}{k_{\mathrm{Lig}}} \qquad (1)$$

where D_{Chd} is the Chordin diffusion coefficient and k_{Lig} is the binding rate of the BMP ligand to Chordin. This profile is valid in most places, ($x > \varepsilon$ with $\varepsilon \propto 1/[\mathrm{BMP}]_{\mathrm{tot}} \to 0$). It is robust to changes in the levels of network components but does not scale with embryo size; indeed, embryo size does not appear in equation (1), and thus does not influence the shape of the gradient. Similarly, solving the model numerically under conditions of secondary-axis induction, we find that two axes ensue, but these two axes decay at the same rate as the original axis and do not scale to half-embryo size (Fig. 2a–c).

Thus, shuttling of a single ligand is not sufficient for scaling. We extended the model to account for the additional ligand Admp, and its feedback-mediated repression. This model can also be solved analytically (Box 1), predicting the activation profile:

$$S(x) \approx \frac{T_{\mathrm{Admp}}}{(x/L)^2} \qquad (2)$$

where T_{Admp} denotes the BMP concentration threshold at which *admp* is repressed. Again, this profile is valid in most positions x. This activation profile has two important consequences. First, the shape of the profile depends only on T_{Admp} and is independent of the other parameters in the system. Accordingly, the profile is robust to fluctuations in most parameters (Fig. 2g). Experimental support for this robustness is provided by the fact that depletion of *bmp2*, *bmp4* or *bmp7*, as well as the partial depletion of Chordin using antisense morpholino directed against one of the *chordin* pseudo alleles, display only a minor phenotype[35,36].

A second notable feature of the activation profile equation (2) is the explicit scaling of position (x) with embryo length (L). In fact, the activation profile is a function of the ratio x/L, implying the scaling of pattern with size. For example, a gene that is normally induced at 50% embryo length ($x/L = \frac{1}{2}$) will be expressed at 50% embryo length irrespective of embryo size, and in particular will be found in the middle of a dorsal-half embryo (Fig. 2h). Moreover, solving the model under conditions of secondary-axis induction, with the addition of ventral source of secreted inhibitor, we find the two axes are now properly scaled (Fig. 2i). Indeed, this scaling of the twinned embryo can be readily explained by equation (2): when two sources of inhibitor are present at opposite poles, the symmetry positions the new ventral side at the mid-point ($x = L/2$), and each gradient is now defined separately with respect to this point. The resulting twinned gradients will thus be given by equation (2), with an effective embryo length of $L/2$.

Experimental evidence for shuttling

The shuttling mechanism predicts three molecular features that are required for this mechanism. First, the binding of Chordin to Bmp is predicted to be of higher affinity than its binding to Admp, an assessment that is supported by the relatively high concentrations of Chordin required for Admp inhibition[22]. Second, cleavage of Chordin is predicted to occur primarily when in complex with a BMP ligand. Chordin is in fact degraded by Xlr also in the absence of BMP, at least *in vitro*[21]. We propose that the formation of an xTsg–BMP complex, which facilitates both the binding of BMPs to Chordin and the degradation of the tertiary complex Chordin–xTsg–BMP, accounts for this assumption.

The third, key prediction of our model is the shuttling of Bmp and Admp by Chordin away from their domain of production. Shuttling requires that the BMP ligands diffuse primarily when bound to Chordin. An attenuated (effective) diffusion of free BMP can be achieved by various mechanisms, including binding to immobilized receptors or elements of the extracellular matrix[37], rapid degradation of receptor-bound ligand[38] or excessively high abundance of Chordin[39] (Supplementary Information). Although the diffusion of the BMPs was not assayed directly, it was shown that in animal caps, which do not express Chordin, BMP4 functions as short-range ligand[40].

To examine the predicted shuttling, we injected messenger RNA coding for Myc-tagged BMP4 into the dorsal part of early embryos, and followed its distribution by direct immunostaining at a later stage. The original injection site was marked by the co-injection of mRNA encoding cytoplasmic green fluorescent protein (GFP). In a parallel experiment, we injected the Myc-tagged *bmp4* mRNA with *chordin* morpholino, thus depleting Chordin, the presumed shuttling molecule. As predicted, we found that BMP4–Myc is shuttled to the

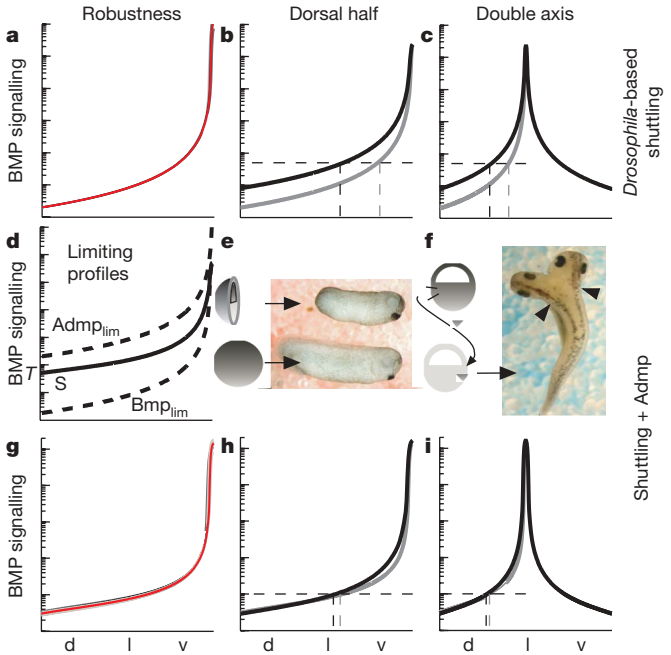


Figure 2 | Shuttling model supports size compensation and secondary-axis induction. **a–c**, 'Core' shuttling model. The model consists of a single BMP ligand that is not subject to auto-repression. **a**, Robustness to twofold decrease (dashed line) or increase (full line) in the production rate of BMP (light grey), Chordin (dark grey) and the protease (black). Unperturbed system profile in red. All profiles are virtually overlapping. **b**, BMP activation profile in wild type (grey) and dorsal-half embryo (black). **c**, BMP activation profile in wild-type embryo (grey) and in embryos with an added ventral source of inhibitor (black). The x axes in **b** and **c** are scaled to embryo length. **d–i**, *Xenopus*-based shuttling. The 'core' shuttling model was extended to include also Admp and its auto-repression by the BMP pathway (Box 1). This model now includes two ligands (Bmp and Admp), both of which can bind Chordin, albeit with a different affinity. d, l, v, dorsal, lateral and ventral, respectively. **d**, Activation profile of *Xenopus*-based shuttling. Each BMP ligand can be associated with a 'limiting profile', based on its predicted behaviour by the *core* shuttling mechanism, denoted by $\mathrm{Admp}_{\mathrm{lim}}$ and $\mathrm{Bmp}_{\mathrm{lim}}$ (dashed lines). The actual signalling profile, S (black), is a weighted average of the two limiting profiles, and can obtain any intermediate value, depending on the relative levels of Admp and Bmp (Box 1). Because Admp is subject to auto-repression, its level at the dorsal-most region is set to the repression threshold, T. This is possible only when $\mathrm{Admp}_{\mathrm{lim}} > \mathrm{Bmp}_{\mathrm{lim}}$, implying a tighter binding of Chordin to Bmp than to Admp. **e**, Scaling of dorsal-half embryos. A dorsal-half embryo in tadpole stage compared with a sibling wild-type embryo. **f**, Double axis. Dorsal view of an embryo with a double axis. Arrowheads point to the two axes. **g–i**, Robustness and scaling. **g**, The extended shuttling model maintains its robustness for ligand distribution. **h**, In addition, it supports the precise scaling of pattern with size. **i**, The extended model ensures precise axis duplication upon ventral transplantation of an organizer. Medium grey in **g** denotes double or half Admp production rate (full/dashed). Other line colours in **g–i** are as in **a–c**. The x axis in **h–i** is scaled to embryo length. See Supplementary Information for the parameters used.

Box 1 | 'Shuttling mechanism' in the *Xenopus* network: model formation

Our model is based on the shuttling mechanism[26]. The 'core' shuttling model previously described[26] considers a single BMP ligand whose total levels are fixed in time. Here, we extend this model to include two BMP ligands: Admp and Bmp, and to account for the auto-repression of *admp* by the BMP pathway activation. This model is obtained as a limiting case of the more general set of equations used in our numerical screen (equation (3)), by assuming that the free ligands do not diffuse ($D_{Lig} = 0$), and that the free inhibitor is not degraded ($\lambda_{Chd} = 0$). We are interested in the activation profile, $S(x) = [Bmp](x) + [Admp](x)$, as measured at steady state. This can be derived by solving the following set of equations:

$$0 = D_{Chd}\nabla^2[Chd] - k_{Admp}[Chd][Admp] - k_{Bmp}[Chd][Bmp]$$

$$0 = -k_{Admp}[Chd][Admp] + \lambda_{Chd}^{Admp}[XIr][ChdAdmp]$$

$$0 = -k_{Bmp}[Chd][Bmp] + \lambda_{Chd}^{Bmp}[XIr][ChdBmp] \quad (4)$$

$$0 = D_{Comp}\nabla^2[ChdAdmp] + k_{Admp}[Chd][Admp] - \lambda_{Chd}^{Admp}[XIr][ChdAdmp]$$

$$0 = D_{Comp}\nabla^2[ChdBmp] + k_{Bmp}[Chd][Bmp] - \lambda_{Chd}^{Bmp}[XIr][ChdBmp]$$

We consider one-dimensional geometry (Supplementary Information), with the dorsal-most region at $x = L$ and the ventral-most at $x = 0$. We assume that all fluxes of the diffusing quantities ([*Chd*], [*ChdAdmp*] and [*ChdBmp*]) vanish at $x = 0$. At ($x = L$), [*Chd*] is produced with a constant flux $\left(D_{Chd}\frac{d[Chd]}{dx}_{x=L} = \eta_{Chd}\right)$, whereas the fluxes of the complexes are zero.

Defining $S_0^{Lig} \equiv 2D_{Chd}/k_{Lig}$, ligand being Admp or Bmp, the solution to equation (4) is given by (see Supplementary Information, section 3):

$$Admp(x) = (1-\delta)\frac{S_0^{Admp}}{x^2+\varepsilon^2}; Bmp(x) = \delta\frac{S_0^{Bmp}}{x^2+\varepsilon^2} \quad (5)$$

where $0 < \delta < 1$ is a constant that depends on the relative levels of total Admp versus total Bmp

$$\delta = \frac{k_{Bmp}Bmp^{tot}}{k_{Bmp}Bmp^{tot} + k_{Admp}Admp^{tot}} \quad (6)$$

and ε is an integration coefficient whose level is inversely proportional to the average total level of the two BMP ligands in the system. The robust scaling solution is obtained when this level is sufficiently large, so that $\varepsilon \ll L$. In this case, the signalling level in most places in the embryo is given by:

$$S(x) \approx \frac{S_0}{x^2}; S_0 = (1-\delta)S_0^{Admp} + \delta S_0^{Bmp} \quad (7)$$

As noted above, δ, and accordingly S_0, depends on the relative level of Admp versus Bmp. Thus, a range of solutions is possible. In fact, depending on δ, the solutions can lie anywhere between the limiting solution corresponding to the case where only Bmp is present ($\delta = 1$) to the limiting solution obtained when only Admp is present ($\delta = 0$). This provides the system with the flexibility to define the precise solution by self-regulating the levels of Admp and Bmp and thus of δ. The scaled solution is obtained when the following three conditions are satisfied. First, Chordin binds Bmp with a higher affinity than Admp ($k_{Bmp} > k_{Admp}$) as we observed numerically (Fig. 1i, j). The lower affinity of Chordin to Admp allows for the accumulation of free Admp; consequently, the limiting profile corresponding to Bmp only is lower than the profile corresponding to Admp only (Fig. 2d). Second, Bmp levels are relatively fixed in time (its production and degradation are balanced), but Admp continues to accumulate as long as it is produced (exhibits a slow degradation). Third, Admp is repressed by BMP signalling, and the repression threshold T_{Admp} allows for its full repression throughout the embryo at some intermediate profile $0 < \delta < 1$. Steady state will be achieved when Admp has accumulated to just the right level to repress its own production everywhere, and in particular at the dorsal-most pole, where signalling level is lowest. Note that we assume that the level of free ligand is sufficient for signalling in the presence of high levels of Chordin. Thus, steady state is obtained when $S(x = L) = T_{Admp}$. Substituting this condition into the signalling profile, equation (7), we obtain the scaled solution (equation (2) in main text):

$$S(x) \approx \frac{T_{Admp}}{(x/L)^2} \quad (8)$$

which is valid for most positions of the embryo that satisfy $x > \varepsilon$. In the Supplementary Information we show that this same activation profile is indeed obtained as a numerical solution of the full dynamics, and can be realized within an extended model that includes additional known components of the patterning network.

ventral pole in embryos that expressed Chordin (Fig. 3b–e), but remains tightly localized to the dorsal pole in the embryos that were depleted of Chordin (Fig. 3f–i). This experiment thus provides direct evidence for the shuttling of BMP4 away from its site of production, and verifies that Chordin is required for this shuttling.

To verify further that shuttling is used during the patterning process itself, we examined the expression domains of ventral genes during secondary-axis induction. Ventral injection of *siamois* (*sia*)

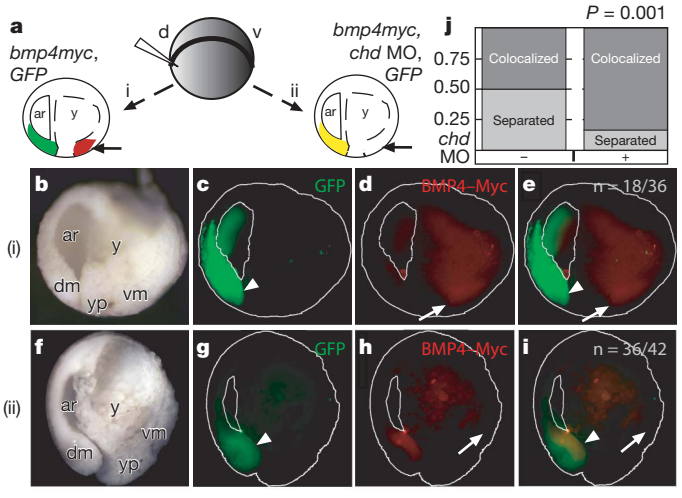
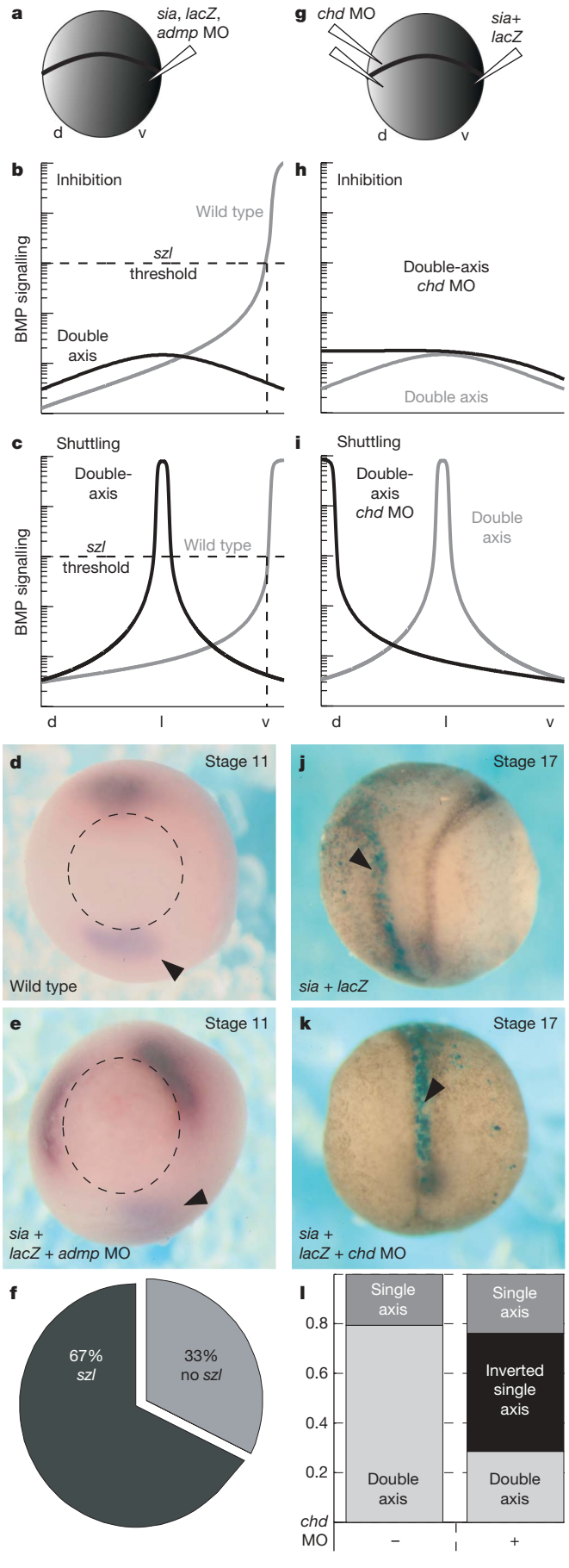


Figure 3 | Direct visualization of BMP4 shuttling by Chordin. **a**, Experimental design. Embryos at the two-cell stage were co-injected on the dorsal side with *bmp4myc* (400 pg) and *GFP* (260 pg) mRNAs, with or without *chordin* morpholino (250 μM). Embryos were grown to stage 12 and analysed by immunohistochemistry. Completely ventralized embryos were not analysed. The expected outcome in the case of shuttling is shown in scheme (i), whereas scheme (ii) displays the expected outcome in the absence of shuttling. Owing to the presence of the yolk in the middle of the embryo, we expect to observe the shuttled BMP4–Myc only in the region indicated by the arrow. ar, archenteron; y, yolk; *chd* MO, *chordin* morpholino. **b**–**e**, Shuttling in embryos co-injected with *GFP* and *bmp4myc* mRNA. Stage 12 embryos injected with *bmp4myc* and *GFP* were sectioned along the dorsoventral axis. **b**, Embryo morphology. dm, dorsal mesoderm; yp, yolk-plug; vm, ventral mesoderm. **c**, GFP fluorescence (green) marking the lineage of cells injected with *bmp4myc*. The dorsal mesoderm and dorsal lip are clearly visible (arrowhead). **d**, Distribution of BMP4–Myc. **e**, Merge of **c** and **d**. BMP4–Myc is concentrated at the ventral mesoderm, marked by an arrow, and is absent from the injection site (arrowhead). Auto-fluorescence in the yolk region was detected also in uninjected controls, but was not seen in the mesoderm (not shown). **f**–**i**, Inhibition of shuttling by *chordin* morpholino. Stage 12 embryos injected with *bmp4myc*, *chordin* morpholino and *GFP* were sectioned along the dorsoventral axis. **f**, Embryo morphology. **g**, GFP (green) marking the lineage of the injected cells. The dorsal mesoderm and dorsal lip are clearly visible (arrowhead). **h**, Distribution of BMP4–Myc. **i**, Merge of **g** and **h**. BMP4–Myc is absent from the ventral mesoderm (arrow) but remained at the site of injection (arrowhead). **j**, Statistics. In 50% of embryos injected with *GFP* and *bmp4myc*, BMP4–Myc was separated from the GFP lineage tracer, as predicted by the shuttling model ($n = 36$). Eighty-three per cent of the embryos injected also with *chordin* morpholino showed co-localization of GFP and BMP4–Myc, indicating no shuttling ($n = 42$). $P = 0.001$ (Fisher's exact test).



mRNA[41] leads to a secondary axis, with duplicated expression of all dorsal genes[42,43]. We reasoned that the expression domains of the *ventral* genes can provide evidence for shuttling. Consider ventral co-injection of *sia* with *admp* morpholino (Fig. 4a): in such an experiment, the induced organizer serves merely as an additional source of inhibitors. Expression of ventral genes in the region between the two organizers, which corresponds to the lateral region of the wild-type embryo, will occur if the BMP signalling level is increased above its normal level. This will happen only if BMP ligands are concentrated in this region; that is, if the new organizer causes BMP shuttling (Fig. 4b, c). We indeed observed the expression of the ventral marker, *sizzled*, between the two organizers, supporting the shuttling of BMP ligands by components of the new organizer (Fig. 4d–f).

As an additional assay, we attempted to invert the embryonic axis by concomitantly inducing an organizer in the prospective ventral

Figure 4 | Experimental evidence for shuttling in *Xenopus*. **a**–**f**, Expression of ventral genes in double-axis embryos: **a**, Experimental design. Embryos at the two-cell stage were co-injected on the ventral side with *sia* mRNA (27 pg) to induce a secondary organizer, *admp* morpholino (MO; 100 μM) to inhibit the translation of *admp* in the newly induced organizer, and *lacZ* expression plasmid for lineage tracing (see Methods). The injected embryos were grown to stage 11, when expression of *chordin* and *sizzled* was monitored by *in situ* hybridization. **b**, Prediction of inhibition-based model. We solved numerically for the predicted BMP activation profile in double-injected embryos. Because the co-injection of *sia* mRNA and *admp* morpholino culminates in the expression of BMP inhibitors such as Noggin and Chordin (but not BMP ligands) in the induced organizer, the inhibition-based model predicts an overall reduction in the level of BMP activation, and consequently the loss of expression of ventral genes, such as *sizzled*. Grey, wild-type, black-manipulated embryo. d, l, v, dorsal, lateral and ventral, respectively. **c**, Prediction of shuttling-based model. In contrast to the inhibition-based model, this model predicts a re-distribution of BMP, with a significantly elevated level at mid-embryo, allowing the expression of ventral genes. Grey, wild-type, black-manipulated embryo. **d**, Expression of *sizzled* in wild-type embryo. Vegetal view (dorsal to the top) of a stage 11 wild-type embryo stained for *sizzled* (light purple, arrowhead). The dorsal region is marked by *chordin* expression (magenta). Dotted circle marks the blastopore. **e**, Expression of *sizzled* in injected embryo. Embryo orientation as in **d**. β-gal activity was detected in the dorsal–animal part (not shown). *chordin* mRNA is expressed in both organizers, and expression of *sizzled* (arrowhead) is observed between the two sites of *chordin* expression, despite the increased expression of inhibitors. **f**, Fraction of double-injected embryos showing *sizzled* expression as in **e** ($n = 54$). **g**–**l**, Axis inversioning. The experimental design was as follows. Embryos at the two-cell stage were co-injected with *sia* (27 pg) and *lacZ* expression plasmid on the ventral side and *chordin* morpholino (250 μM, double injection) on the dorsal side. The injected embryos were grown to stage 17 and assayed for double-axis formation by morphology. **h**, Prediction of inhibition-based model. We solved numerically for the predicted BMP activation profile in double-injected embryos (Supplementary Information). Within the inhibition-based model, the double-axis observed upon ventral injection of *sia* (grey) is maintained also upon the dorsal co-injection of *chordin* morpholino (black). **i**, Prediction of shuttling-based model. In contrast to the inhibition-based model, the re-distribution of BMP predicted by the shuttling model will lead to a single, inverted axis, with maximal BMP activity at the original dorsal side (black); *sia*-induced double axis in grey. **j**, Double axis in embryo injected with *sia*. Dorsal view of stage 17 embryos; anterior is to the top. Embryos were tested for β-gal expression with X-gal (blue). The induced axis is stained (arrowhead), whereas the original axis is not. **k**, Single inverted axis in embryo co-injected with *sia* and *chordin* morpholino. The single axis expresses β-gal (blue), indicating that it is derived from the induced organizer. Orientation as in **j**. **l**, Fraction of double-injected embryos showing inverted axis. *sia* mRNA induced a secondary axis in 79% of embryos (light grey, $n = 29$). When co-injected with *chordin* morpholino at the dorsal side, a double axis was induced in only 28% of the embryos ($n = 21$). Twenty-four per cent of the embryos had a single axis with no β-gal staining (dark grey), corresponding to embryos where a double axis was not induced (compared with 21% when *sia* was injected without *chordin* morpholino); 48% had a single axis with β-gal staining, corresponding to an axis induced by *sia* mRNA injection (black).

side (through the injection of *sia*) and depleting Chordin from the original organizer (Fig. 4g). Depletion of *chordin* from wild-type embryos has a mild effect, and does not abolish the dorsoventral polarity (probably because of the presence of additional BMP inhibitors at the organizer[15,35], not shown). Accordingly, in the absence of shuttling, the predicted outcome of co-inducing a second organizer while inhibiting Chordin at the original organizer is a double axis, with the original dorsal axis somewhat ventralized (Fig. 4h). In contrast, axis inversion is expected if the induced organizer leads to shuttling of BMP ligands to the original dorsal side, because high BMP signalling will repress the expression of dorsally expressed BMP inhibitors (Fig. 4i). Axis inversion was indeed observed, providing additional support for shuttling (Fig. 4j–l). We note that this experimental outcome is in agreement with the observation that dorsal lip grafts seldom induce secondary axes when prepared from embryos injected with *chordin* morpholino[35], a result that provides additional support for the shuttling mechanism.

Conclusions

We provide a simple, quantitative explanation for the capacity of the *Xenopus* embryo to scale pattern with size. Three key features of the BMP patterning network underlie this capacity. First, patterning is governed by a shuttling-based mechanism, where the BMP ligands are effectively transported by a common BMP inhibitor (Chordin) to the ventral-most part of the embryo, establishing a sharp, power-law decaying activation profile. Second, the presence of two BMP ligands, which differ in their affinity to the inhibitor Chordin, allows for a range of possible steady-state profiles, depending on the relative abundance of the two ligands (Box 1). This is in contrast to the case of a single ligand, where the gradient approaches a unique 'limiting profile' independent of the total ligand level. Finally, the negative auto-repression of the BMP ligand, Admp, is used for sensing embryo size, and effectively tunes the pattern with size. Together, these three features lead to a robust and sharp gradient that is properly scaled with embryo size.

The model we describe is based on what appears to be the core of the patterning network. It can be extended, however, to include additional network components, such as production of Chordin over a large field and the regulation of Chordin expression, xTsg expression[44] and the BMP receptors (Supplementary Information). Further analysis is required to characterize the quantitative contributions of the added network components to patterning, in particular their possible impact on the dynamics of gradient formation.

Our work predicts an evolutionary course that endowed the BMP signalling pathway with two properties that are seemingly mutually exclusive: robust patterning and the ability scale pattern with size. Using the shuttling capacity of the inhibitors with ligands that display similar properties provides robustness but excludes scaling. The evolution of a ligand with unique regulatory properties provides the added feature of scaling, without compromising robustness.

METHODS SUMMARY

Numerical screen. We consider two BMP ligands, [Bmp] and [Admp], an inhibitor, Chordin ([Chd]), the respective complexes [ChdBmp] and [ChdAdmp] and a protease [Xlr]. The model is defined by the following set of reaction diffusion equations:

$$\begin{aligned}
\frac{\partial[\mathrm{Chd}]}{\partial t} &= D_{\mathrm{Chd}}\nabla^2[\mathrm{Chd}] - k_{\mathrm{Admp}}[\mathrm{Admp}][\mathrm{Chd}] - k_{\mathrm{Bmp}}[\mathrm{Bmp}][\mathrm{Chd}] - \lambda_{\mathrm{Chd}}[\mathrm{Xlr}][\mathrm{Chd}] \\
\frac{\partial[\mathrm{Admp}]}{\partial t} &= D_{\mathrm{Lig}}\nabla^2[\mathrm{Admp}] - k_{\mathrm{Admp}}[\mathrm{Admp}][\mathrm{Chd}] + \lambda_{\mathrm{Chd}}^{\mathrm{Admp}}[\mathrm{Xlr}][\mathrm{ChdAdmp}] \\
\frac{\partial[\mathrm{Bmp}]}{\partial t} &= D_{\mathrm{Lig}}\nabla^2[\mathrm{Bmp}] - k_{\mathrm{Bmp}}[\mathrm{Chd}][\mathrm{Bmp}] + \lambda_{\mathrm{Chd}}^{\mathrm{Bmp}}[\mathrm{Xlr}][\mathrm{ChdBmp}] \qquad (3) \\
\frac{\partial[\mathrm{ChdAdmp}]}{\partial t} &= D_{\mathrm{Comp}}\nabla^2[\mathrm{ChdAdmp}] + k_{\mathrm{Admp}}[\mathrm{Admp}][\mathrm{Chd}] - \lambda_{\mathrm{Chd}}^{\mathrm{Admp}}[\mathrm{Xlr}][\mathrm{ChdAdmp}] \\
\frac{\partial[\mathrm{ChdBmp}]}{\partial t} &= D_{\mathrm{Comp}}\nabla^2[\mathrm{ChdBmp}] + k_{\mathrm{Bmp}}[\mathrm{Bmp}][\mathrm{Chd}] - \lambda_{\mathrm{Chd}}^{\mathrm{Bmp}}[\mathrm{Xlr}][\mathrm{ChdBmp}]
\end{aligned}$$

The steady-state signalling profile $S(x) = [\mathrm{Admp}](x) + [\mathrm{Bmp}](x)$ is considered as the biologically relevant output.

Boundary conditions. All fluxes vanish at $x = 0$ and $x = L$ except a constant flux of Chordin, and signal-dependent flux of Admp at the dorsal pole ($x = L$).

Screen parameters. The following nine parameters were changed in the screen: diffusion coefficients, binding of Chordin with Bmp or Admp, cleavage of Chordin by Xlr when it is free or in complex with Bmp or Admp, and the Chordin flux. The respective parameter space was scanned systematically with each parameter modified by at least two orders of magnitude, and the model was solved numerically for over 26,000 parameter sets. A network was marked 'consistent' if the associated signalling profile displayed proper dorsoventral polarity, scaled with embryo size, and was robust to parameter variations (Supplementary Information).

The shuttling coefficient S_h measures the dynamic range of the total ligand, normalized by the dynamic range of the total free ligand. For ideal shuttling $S_h \approx 1$, whereas for the inhibitory model $S_h \approx 0$.

Embryo manipulation. Embryos at the desired stage were fixed in MEMFA and processed for *in situ* hybridization. Antisense morpholino oligonucleotides and mRNA were injected to embryos at the two-cell stage. β-Galactosidase (β-gal) and *GFP*[45] activities were used for lineage tracing. Embryos were injected in $1 \times$ MBSH, and raised to the desired stage in $0.1 \times$ MBSH.

Immunohistochemistry. Embryos were fixed, re-hydrated and bisected along the dorsoventral axis. After blocking, embryos were incubated with primary antibody overnight, and then labelled with a fluorescent secondary antibody.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 27 January; accepted 8 May 2008.

1. Waddington, C. H. Canalization of development and the inheritance of acquired characters. *Nature* **150,** 563–565 (1942).
2. Kirschner, M. & Gerhart, J. Evolvability. *Proc. Natl Acad. Sci. USA* **95,** 8420–8427 (1998).
3. De Robertis, E. M. Spemann's organizer and self-regulation in amphibian embryos. *Nature Rev. Mol. Cell Biol.* **7,** 296–302 (2006).
4. Spemann, H. *Embryonic Development and Induction* (Yale Univ. Press, New Haven, 1938).
5. Spemann, H. & Mangold, H. Induction of embryonic primordia by implantation of organizers from a different species. *Roux's Arch. Entw. Mech.* **100,** 599–638 (1924).
6. Harland, R. M. Neural induction in *Xenopus*. *Curr. Opin. Genet. Dev.* **4,** 543–549 (1994).
7. Cooke, J. Scale of body pattern adjusts to available cell number in amphibian embryos. *Nature* **290,** 775–778 (1981).
8. Lowe, C. J. *et al.* Dorsoventral patterning in hemichordates: insights into early chordate evolution. *PLoS Biol.* **4,** e291 (2006).
9. Marques, G. *et al.* Production of a DPP activity gradient in the early *Drosophila* embryo through the opposing actions of the SOG and TLD proteins. *Cell* **91,** 417–426 (1997).
10. Ferguson, E. L. Conservation of dorsal–ventral patterning in arthropods and chordates. *Curr. Opin. Genet. Dev.* **6,** 424–431 (1996).
11. De Robertis, E. M. & Kuroda, H. Dorsal-ventral patterning and neural induction in *Xenopus* embryos. *Annu. Rev. Cell Dev. Biol.* **20,** 285–308 (2004).
12. Francois, V. & Bier, E. *Xenopus* chordin and *Drosophila* short gastrulation genes encode homologous proteins functioning in dorsal–ventral axis formation. *Cell* **80,** 19–20 (1995).
13. Sasai, Y. *et al. Xenopus* chordin: a novel dorsalizing factor activated by organizer-specific homeobox genes. *Cell* **79,** 779–790 (1994).
14. Sasai, Y., Lu, B., Steinbeisser, H. & De Robertis, E. M. Regulation of neural induction by the Chd and Bmp-4 antagonistic patterning signals in *Xenopus*. *Nature* **376,** 333–336 (1995).
15. Khokha, M. K., Yeh, J., Grammer, T. C. & Harland, R. M. Depletion of three BMP antagonists from Spemann's organizer leads to a catastrophic loss of dorsal structures. *Dev. Cell* **8,** 401–411 (2005).
16. Ross, J. J. *et al.* Twisted gastrulation is a conserved extracellular BMP antagonist. *Nature* **410,** 479–483 (2001).
17. Oelgeschlager, M., Larrain, J., Geissert, D. & De Robertis, E. M. The evolutionarily conserved BMP-binding protein Twisted gastrulation promotes BMP signalling. *Nature* **405,** 757–763 (2000).
18. Chang, C. *et al.* Twisted gastrulation can function as a BMP antagonist. *Nature* **410,** 483–487 (2001).
19. Larrain, J. *et al.* Proteolytic cleavage of Chordin as a switch for the dual activities of Twisted gastrulation in BMP signaling. *Development* **128,** 4439–4447 (2001).
20. Goodman, S. A. *et al.* BMP1-related metalloproteinases promote the development of ventral mesoderm in early *Xenopus* embryos. *Dev. Biol.* **195,** 144–157 (1998).
21. Piccolo, S. *et al.* Cleavage of Chordin by Xolloid metalloprotease suggests a role for proteolytic processing in the regulation of Spemann organizer activity. *Cell* **91,** 407–416 (1997).
22. Dosch, R. & Niehrs, C. Requirement for anti-dorsalizing morphogenetic protein in organizer patterning. *Mech. Dev.* **90,** 195–203 (2000).

23. Moos, M. Jr, Wang, S. & Krinks, M. Anti-dorsalizing morphogenetic protein is a novel TGF-beta homolog expressed in the Spemann organizer. *Development* **121,** 4293–4301 (1995).
24. Reversade, B. & De Robertis, E. M. Regulation of ADMP and BMP2/4/7 at opposite embryonic poles generates a self-regulating morphogenetic field. *Cell* **123,** 1147–1160 (2005).
25. Decotto, E. & Ferguson, E. L. A positive role for Short gastrulation in modulating BMP signaling during dorsoventral patterning in the *Drosophila* embryo. *Development* **128,** 3831–3841 (2001).
26. Eldar, A. *et al.* Robustness of the BMP morphogen gradient in *Drosophila* embryonic patterning. *Nature* **419,** 304–308 (2002).
27. Umulis, D. M., Serpe, M., O'Connor, M. B. & Othmer, H. G. Robust, bistable patterning of the dorsal surface of the *Drosophila* embryo. *Proc. Natl Acad. Sci. USA* **103,** 11613–11618 (2006).
28. Mizutani, C. M. *et al.* Formation of the BMP activity gradient in the *Drosophila* embryo. *Dev. Cell* **8,** 915–924 (2005).
29. Meinhardt, H. & Roth, S. Developmental biology: sharp peaks from shallow sources. *Nature* **419,** 261–262 (2002).
30. Shimmi, O., Umulis, D., Othmer, H. & O'Connor, M. B. Facilitated transport of a Dpp/Scw heterodimer by Sog/Tsg leads to robust patterning of the *Drosophila* blastoderm embryo. *Cell* **120,** 873–886 (2005).
31. Wang, Y. C. & Ferguson, E. L. Spatial bistability of Dpp-receptor interactions during *Drosophila* dorsal–ventral patterning. *Nature* **434,** 229–234 (2005).
32. van der Zee, M., Stockhammer, O., von Levetzow, C., da Fonseca, R. N. & Roth, S. Sog/Chordin is required for ventral-to-dorsal Dpp/BMP transport and head formation in a short germ insect. *Proc. Natl Acad. Sci. USA* **103,** 16307–16312 (2006).
33. O'Connor, M. B., Umulis, D., Othmer, H. G. & Blair, S. S. Shaping BMP morphogen gradients in the *Drosophila* embryo and pupal wing. *Development* **133,** 183–193 (2006).
34. Eldar, A., Shilo, B. Z. & Barkai, N. Elucidating mechanisms underlying robustness of morphogen gradients. *Curr. Opin. Genet. Dev.* **14,** 435–439 (2004).
35. Oelgeschlager, M., Kuroda, H., Reversade, B. & De Robertis, E. M. Chordin is required for the Spemann organizer transplantation phenomenon in *Xenopus* embryos. *Dev. Cell* **4,** 219–230 (2003).
36. Reversade, B., Kuroda, H., Lee, H., Mays, A. & De Robertis, E. M. Depletion of Bmp2, Bmp4, Bmp7 and Spemann organizer signals induces massive brain formation in *Xenopus* embryos. *Development* **132,** 3381–3392 (2005).
37. Ohkawara, B., Iemura, S.-I., ten Dijke, P. & Ueno, N. Action range of BMP is defined by its N-terminal basic amino acid core. *Curr. Biol.* **12,** 205–209 (2002).
38. Lander, A. D., Nie, Q. & Wan, F. Y. Do morphogen gradients arise by diffusion? *Dev. Cell* **2,** 785–796 (2002).
39. Lee, H. X., Ambrosio, A. L., Reversade, B. & De Robertis, E. M. Embryonic dorsal–ventral signaling: secreted frizzled-related proteins as inhibitors of tolloid proteinases. *Cell* **124,** 147–159 (2006).
40. Jones, C. M., Armes, N. & Smith, J. C. Signalling by TGF-β family members: short-range effects of Xnr-2 and BMP-4 contrast with the long-range effects of activin. *Curr. Biol.* **6,** 1468–1475 (1996).
41. Lemaire, P., Garrett, N. & Gurdon, J. B. Expression cloning of *Siamois*, a *Xenopus* homeobox gene expressed in dorsal–vegetal cells of blastulae and able to induce a complete secondary axis. *Cell* **81,** 85–94 (1995).
42. Marom, K., Levy, V., Pillemer, G. & Fainsod, A. Temporal analysis of the early BMP functions identifies distinct anti-organizer and mesoderm patterning phases. *Dev. Biol.* **282,** 442–454 (2005).
43. Fagotto, F., Guger, K. & Gumbiner, B. Induction of the primary dorsalizing center in *Xenopus* by the Wnt/GSK/beta-catenin signaling pathway, but not by Vg1, Activin or Noggin. *Development* **124,** 453–460 (1997).
44. Collavin, L. & Kirschner, M. W. The secreted Frizzled-related protein Sizzled functions as a negative feedback regulator of extreme ventral mesoderm. *Development* **130,** 805–816 (2003).
45. Zernicka-Goetz, M. *et al.* An indelible lineage marker for *Xenopus* using a mutated green fluorescent protein. *Development* **122,** 3719–3724 (1996).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Christian for the BMP4 constructs, and the members of our groups for discussions and help with the experiments. This work was supported by Minerva, the Israel Science Foundation and the Hellen and Martin Kimmel award for innovative investigations to N.B. and a grant from the Israel Science Foundation and the Wolfson Family Chair in Genetics to A.F. B-Z.S. holds the Hilda and Cecil Lewis Professorial chair in Molecular Genetics.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to N.B. (Naama.barkai@weizmann.ac.il) or A.F. (fainsod@cc.huji.ac.il).

METHODS

Numerical screen. Full details of the numerical screen are provided in the Supplementary Information (Section 1). Here we summarize the basic methods. Our model does not consider the mechanism underlying the formation of the organizer, but only the subsequent steps[46]. We consider two BMP ligands, [Bmp] and [Admp], an inhibitor, Chordin ([Chd]), the respective complexes [ChdBmp] and [ChdAdmp] and a protease [Xlr]. The model is defined by the following set of reaction diffusion equations:

$$\begin{aligned}
\frac{\partial[\mathrm{Chd}]}{\partial t} &= D_{\mathrm{Chd}}\nabla^2[\mathrm{Chd}] - k_{\mathrm{Admp}}[\mathrm{Admp}][\mathrm{Chd}] - k_{\mathrm{Bmp}}[\mathrm{Bmp}][\mathrm{Chd}] - \lambda_{\mathrm{Chd}}[\mathrm{Xlr}][\mathrm{Chd}] \\
\frac{\partial[\mathrm{Admp}]}{\partial t} &= D_{\mathrm{Lig}}\nabla^2[\mathrm{Admp}] - k_{\mathrm{Admp}}[\mathrm{Admp}][\mathrm{Chd}] + \lambda_{\mathrm{Chd}}^{\mathrm{Admp}}[\mathrm{Xlr}][\mathrm{ChdAdmp}] \\
\frac{\partial[\mathrm{Bmp}]}{\partial t} &= D_{\mathrm{Lig}}\nabla^2[\mathrm{Bmp}] - k_{\mathrm{Bmp}}[\mathrm{Chd}][\mathrm{Bmp}] + \lambda_{\mathrm{Chd}}^{\mathrm{Bmp}}[\mathrm{Xlr}][\mathrm{ChdBmp}] \qquad (3) \\
\frac{\partial[\mathrm{ChdAdmp}]}{\partial t} &= D_{\mathrm{Comp}}\nabla^2[\mathrm{ChdAdmp}] + k_{\mathrm{Admp}}[\mathrm{Admp}][\mathrm{Chd}] - \lambda_{\mathrm{Chd}}^{\mathrm{Admp}}[\mathrm{Xlr}][\mathrm{ChdAdmp}] \\
\frac{\partial[\mathrm{ChdBmp}]}{\partial t} &= D_{\mathrm{Comp}}\nabla^2[\mathrm{ChdBmp}] + k_{\mathrm{Bmp}}[\mathrm{Bmp}][\mathrm{Chd}] - \lambda_{\mathrm{Chd}}^{\mathrm{Bmp}}[\mathrm{Xlr}][\mathrm{ChdBmp}]
\end{aligned}$$

The steady-state signalling profile $S(x) = [Admp](x) + [Bmp](x)$ is considered as the biologically relevant output.

Boundary conditions. All fluxes vanish at $x = 0$ and $x = L$ except a constant flux of Chordin, η_{Chd}, and signal-dependent flux of Admp, $\alpha(S)$, at the dorsal pole ($x = L$), with $\alpha(S) = 10^{-3}\dfrac{T_{\mathrm{Admp}}^4}{T_{\mathrm{Admp}}^4 + S(L)^4}\,\mu\mathrm{M}\mu\mathrm{m}\mathrm{s}^{-1}$.

Screen parameters. The following parameters were changed in the screen: diffusion coefficients of Chordin, the ligands and the Chordin–ligand complexes (D_{Chd}, D_{Lig}, D_{Comp}), binding of the Chordin with Bmp or Admp (k_{Bmp}, k_{Admp}), cleavage of the Chordin by Xlr when it is free or in complex with Bmp or Admp $\left(\lambda_{\mathrm{Chd}}, \lambda_{\mathrm{Chd}}^{\mathrm{Bmp}}, \lambda_{\mathrm{Chd}}^{\mathrm{Admp}}\right)$ and the Chordin flux (η_{Chd}). The respective nine-dimensional parameters space was scanned systematically with each parameter modified by at least two orders of magnitude. Log mid-values of the parameters were: $D_{\mathrm{Chd}} = D_{\mathrm{Lig}} = D_{\mathrm{Comp}} = 1\,\mu\mathrm{m}^2\,\mathrm{s}^{-1}$, $k_{\mathrm{Bmp}} = k_{\mathrm{Admp}} = 0.1\,\mu\mathrm{M}^{-1}\,\mathrm{s}^{-1}$, $\lambda_{\mathrm{Chd}} = \lambda_{\mathrm{Chd}}^{\mathrm{Bmp}} = \lambda_{\mathrm{Chd}}^{\mathrm{Admp}} = 0.1\mu\mathrm{M}^{-1}\mathrm{s}^{-1}$, $\eta_{\mathrm{Chd}} = 10^{1.5}\,\mu\mathrm{M}\,\mu\mathrm{m}\,\mathrm{s}^{-1}$. Xlr concentration was set to [Xlr] = 0.01 μM and whole embryo length $L = 1{,}000\,\mu\mathrm{m}$.

Screen execution. The model was solved numerically for over 26,000 sets of parameters. A network was marked 'consistent' if the associated signalling profile displayed proper dorsoventral polarity, scaled with embryo size and was robust to parameter variations (see Supplementary Information for precise definitions).

Shuttling coefficient. The shuttling coefficient S_{h} measures to the dynamic range of the total ligand $S^{\mathrm{tot}}(x) = S(x) + [\mathrm{ChdAdmp}](x) + [\mathrm{ChdBmp}](x)$, normalized by the dynamic range of the total free ligand, $S(x) = [\mathrm{Admp}](x) + [\mathrm{Bmp}](x)$.

$S_h = \dfrac{\delta S^{\mathrm{tot}}}{\delta S}$, where $\delta S = \dfrac{\max(S) - \min(S)}{\max(S)}$; max(S) and min(S) refer to the maximal and minimal levels of $S(x)$ over the entire embryo, respectively. δS^{tot} is defined similarly. For ideal shuttling $S_{\mathrm{h}} \approx 1$, whereas for the inhibitory model $S_{\mathrm{h}} \approx 0$.

Embryo manipulation. *Xenopus laevis* were purchased from Xenopus1. Embryos were obtained by *in vitro* fertilization and incubated in 0.1 × modified Barth's solution (MBSH). Dorsal-half embryos: stage 8.5 embryos were dechorionated in 0.3 ×MBSH and cut into a dorsal and ventral halves. Dorsal-half embryos were cultured in fresh 0.3 × MBSH until the desired stage.

***In situ* hybridization and probes.** Embryos at the desired stage were fixed in MEMFA and processed for *in situ* hybridization, as described[47]. Digoxigenin (Dig)-labelled RNA and fluorescein probes were transcribed *in vitro* using the RiboMax kit (Promega), and the Dig. RNA labelling mix (Roche) or Fluorescein RNA labelling mix (Roche), respectively. The probes used in the *in situ* hybridization procedure were *chordin*, the Δ59 clone[13] and *sizzled*[48].

Morpholino oligonucleotides, mRNA injections and lineage tracing. Antisense morpholino oligonucleotides (morpholino) were obtained from Gene Tools LLC for *Xenopus laevis. admp* morpholino[24] and *chordin* morpholino (both pseudoalleles[35]) were injected into embryos at the two-cell stage as described in the text. mRNA for microinjection was prepared using the RiboMax kit (Promega), and adding cap analogue (Roche, Pharmacia) at a ratio of 1:5 (GTP:cap analogue). *sia*[41] mRNA was injected ventrally at the two-cell stage. *GFP*[45] and *bmp4myc*[49] were injected dorsally at the two-cell stage as described in the text. β-Gal activity was used for lineage tracing: a *lacZ* expression plasmid (CMV–LacZ), was injected at 30 ng μl^{-1} as described in the text. Staining of the β-gal activity was done as described[42]. Embryos were injected in 1 × MBSH buffer, and raised to the desired stage in 0.1 × MBSH buffer.

Immunohistochemistry. Embryos were fixed in MEMFA for 1 h and stored in Dent's solution (20% dimethylsulphoxide in methanol) at −20 °C overnight. Rehydrated embryos were bisected along the dorsoventral axis, and blocked in PBT (PBS, 2 mg ml^{-1} BSA, 0.1% Triton X-100) plus 10% normal bovine serum. Embryos were incubated with primary antibody (anti-Myc 9E10, 1:100, 4 °C, overnight). The embryos were then washed five times in PBT and incubated with a cy3 fluorescent-labelled secondary antibody (1:500, 4 °C, overnight).

46. Meinhardt, H. Primary body axes of vertebrates: generation of a near-cartesian coordinate system and the role of Spemann-type organizer. *Dev. Dyn.* **235**, 2907–2919 (2006).
47. Fainsod, A., Steinbeisser, H. & De Robertis, E. M. On the function of BMP-4 in patterning the marginal zone of the *Xenopus* embryo. *EMBO J.* **13**, 5015–5025 (1994).
48. Salic, A. N., Kroll, K. L., Evans, L. M. & Kirschner, M. W. Sizzled: a secreted Xwnt8 antagonist expressed in the ventral marginal zone of *Xenopus* embryos. *Development* **124**, 4739–4748 (1997).
49. Sopory, S., Nelsen, S. M., Degnin, C., Wong, C. & Christian, J. L. Regulation of bone morphogenetic protein-4 activity by sequence elements within the prodomain. *J. Biol. Chem.* **281**, 34021–34031 (2006).

Vol 453|26 June 2008|doi:10.1038/nature07011

LETTERS

The Borealis basin and the origin of the martian crustal dichotomy

Jeffrey C. Andrews-Hanna[1], Maria T. Zuber[1] & W. Bruce Banerdt[2]

The most prominent feature on the surface of Mars is the near-hemispheric dichotomy between the southern highlands and northern lowlands. The root of this dichotomy is a change in crustal thickness along an apparently irregular boundary, which can be traced around the planet, except where it is presumably buried beneath the Tharsis volcanic rise[1,2]. The isostatic compensation of these distinct provinces[2,3] and the ancient population of impact craters buried beneath the young lowlands surface[4] suggest that the dichotomy is one of the most ancient features on the planet[3]. However, the origin of this dichotomy has remained uncertain, with little evidence to distinguish between the suggested causes: a giant impact[5,6] or mantle convection/overturn[7–9]. Here we use the gravity[10] and topography[11] of Mars to constrain the location of the dichotomy boundary beneath Tharsis, taking advantage of the different modes of compensation for Tharsis and the dichotomy to separate their effects. We find that the dichotomy boundary along its entire path around the planet is accurately fitted by an ellipse measuring approximately 10,600 by 8,500 km, centred at 67° N, 208° E. We suggest that the elliptical nature of the crustal dichotomy is most simply explained by a giant impact, representing the largest such structure thus far identified in the Solar System.

The origin of the crustal dichotomy remains one of the most fundamental unanswered questions in the study of Mars. Early workers suggested that it may have been produced by a giant impact[5], but the attempted fit of a circular 'Borealis basin' to the irregular dichotomy boundary proved unsatisfactory[12], leaving large regions of the lowlands unexplained (Fig. 1a). Other studies suggested that a number of impacts could be arranged to reproduce the dichotomy boundary[6], although evidence for the individual basins is lacking[1,12]. Alternatively, an endogenic origin of the dichotomy has been advanced, in which a hemispheric-scale mantle upwelling (spherical harmonic degree-1 mantle convection) led to thermal thinning or volcanic thickening of the crust on one side of the planet[7,8]. Similarly, the overturn of a buoyantly unstable cumulate mantle after the solidification of a global magma ocean could result in a degree-1 crustal structure[9].

Additional information is required to discriminate between the proposed theories, notably the original shape of the dichotomy boundary. The surface of Mars has changed much in the >4 billion years since the dichotomy formed. Massive surficial and intrusive volcanism during construction of the Tharsis rise thickened the crust over ~15% of the planet, burying the dichotomy boundary beneath up to 30 km of basalt[13,14]. The location of the boundary is also unclear in Arabia Terra, with crustal thickness and topography intermediate between the highlands and lowlands[1]. Arabia Terra is separated from both the lowlands and highlands by distinct changes in topography and crustal thickness, one, both, or neither of which may be considered the continuation of the dichotomy boundary.

To clarify the nature of Mars's early crustal structure, we inverted the gravity and topography fields under the assumption of crustal thickness-compensated flexure[14]. The model solutions can be represented as a crust of uniform thickness, to which thickness perturbations are added at the top and bottom surfaces (the load and isostatic root, respectively), combined with flexural displacements to reproduce the gravity and topography. The primordial crust of Mars was probably isostatically compensated, with the surface topography buoyantly supported by a thick isostatic root. Subsequent departures from this isostatic state resulted in deflections that were resisted by membrane and flexural stresses in the elastic lithosphere. Although

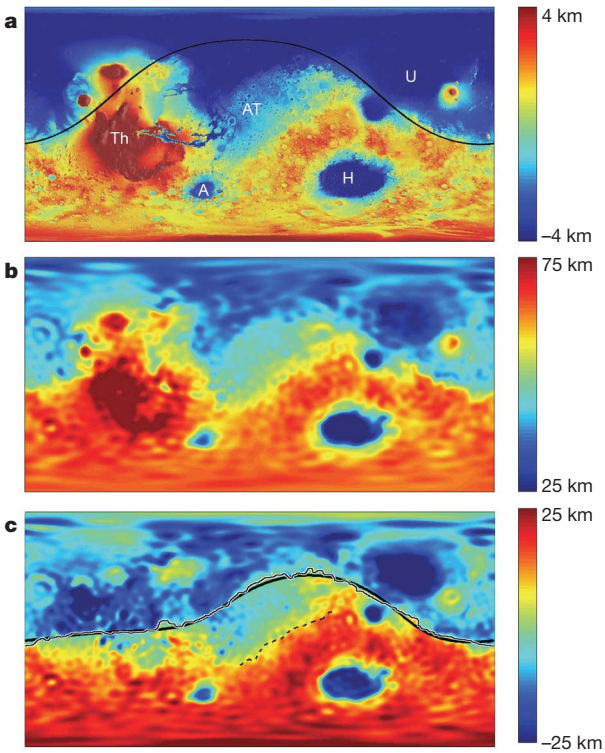


Figure 1 | Topography and crustal structure of Mars. a, Topography[11] and **b**, crustal thickness[2] of Mars (cylindrical projection). Main features labelled in **a** include Tharsis (Th), Arabia Terra (AT), Hellas (H), Argyre (A), and Utopia (U), as well as the Borealis basin outline proposed by Wilhelms and Squyres[5] (solid line). **c**, Modelled bottom crustal thickness perturbation (isostatic root), showing continuation of the dichotomy boundary beneath Tharsis. The observed dichotomy boundary (thin line) is compared with the best-fit ellipse (bold line) in **c**. The break in slope separating Arabia Terra from the highlands is shown as a dashed line.

[1]Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [2]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA.

the observed dichotomy is largely isostatic[2,3] and thus dominated by the isostatic roots, the gravity anomalies[13] and tectonic features[14,15] surrounding Tharsis suggest that the rise is largely a flexurally supported load. Generalizing this thinking, we suggest that the loads and displacements in the Tharsis province should be dominated by Tharsis itself, whereas the isostatic roots should reflect the pre-Tharsis dichotomy, thus enabling us to isolate the dichotomy beneath Tharsis (Supplementary Fig. 1).

The continuation of the dichotomy boundary beneath Tharsis is clearly visible in the model isostatic roots (Fig. 1b). The model cannot account for the full complexity of Tharsis loading, such as a spatially and temporally variable lithosphere thickness during Tharsis construction, or crustal and mantle density anomalies. As a result, the roots are negative beneath younger portions of the rise that are probably supported by a thicker lithosphere (for example Olympus Mons), and positive beneath older portions that are largely isostatic today (for example Tempe Terra). Nevertheless, considering both the present-day topography and the isostatic roots, we are able to map the dichotomy boundary both beneath Tharsis and elsewhere (Fig. 1c; see also Methods). The sub-Tharsis dichotomy is continuous to the west with the observed boundary, and to the east with the break in slope north of Arabia Terra. Although the northern edge of Arabia Terra is sharply defined and continuous with the dichotomy boundary on either side, the southern edge is less distinct and not obviously connected with the boundary (Supplementary Fig. 3c). We also note that the magnetic anomalies distributed throughout much of the highlands are also found within Arabia Terra[16]. Collectively, this evidence suggests that Arabia Terra is a part of the highlands. The cause of its unique topography will be addressed later in this paper.

The path of the dichotomy boundary around the entire planet is now well fitted by an ellipse centred on 67° N, 208° E (Fig. 1c). To best display the shape of the boundary, we re-project Mars in polar coordinates around the best-fit centre of the ellipse (Fig. 2). The ellipse has long and short axes of 10,600 and 8,500 km, respectively, with the long axis oriented towards N76°E, and matches the dichotomy boundary with a root-mean-square misfit of ±100 km (excluding the uncertainty in mapping the boundary location). The full extent of the lowlands now covers 42% of the surface. The elliptical nature of the crustal dichotomy provides a new constraint on models of its formation. Giant impacts with low-angle trajectories produce elliptical basins, such as Hellas on Mars (ratio of major to minor axes, $a/b = 2{,}414\,\text{km}/1{,}820\,\text{km} = 1.33$) and South Pole–Aitken on the Moon ($a/b = 2{,}125\,\text{km}/1{,}542\,\text{km} = 1.38$). Although the Borealis basin far exceeds these and all other impact basins in size, it is comparable in aspect ratio ($a/b = 1.25$). Departures from the elliptical shape can be attributed to local erosion, relaxation and tectonic modification of the boundary[17]. It is unclear at this point whether endogenic processes can explain the elliptical boundary. Although the upwellings in degree-1 convection models range from irregular to quasi-circular in shape[8], there is no known reason to expect a geometric elliptical pattern to emerge.

Global crustal thickness histograms provide a further constraint on the nature of the dichotomy, demonstrating a roughly bimodal distribution[2] between the lowlands and highlands (Fig. 3a). Exclusion of terrains in which subsequent processes have altered the primordial crustal thickness (known basins, Tharsis and Elysium) enhances this bimodal structure, and removal of the anomalous Arabia Terra as well separates the two modes completely. The excavation of crust in a giant impact is known to produce a bimodal distribution of crustal thickness between the basin floor and its surroundings, as demonstrated by Hellas (Fig. 3b). The present-day lowlands crust would have differentiated from an impact-generated local magma ocean. Although there have been no explicit studies of the effects of degree-1 mantle convection on the crustal thickness distribution, convection models are characterized by a centralized upwelling that expands laterally on reaching the base of the lithosphere[7,8]. The resulting gradual decrease in the thermal and mechanical effects away from the upwelling[7] would suggest a similarly gradual decrease in the volcanic thickening or thermal erosion. This is exemplified by Tharsis, which has also been attributed to degree-1 mantle convection[18]. A crustal thickness histogram of the southern portion of Tharsis and a nearly equal area of adjacent highlands (Fig. 3c) demonstrates that in this case degree-1 volcanic thickening did not produce a bimodal distribution.

An impact origin for the crustal dichotomy suggests that other secondary basin-related features should have formed. However, the superposition of a saturated crater population over the dichotomy has erased all but the large-scale crustal thickness signature. Giant impact basins are commonly surrounded by a multiple ring system, with radii scaled by factors of $\sqrt{2}$ (~1.41) relative to the basin radius[19]. It is interesting to note that the southern edge of Arabia Terra roughly parallels the dichotomy boundary at a mean distance from the ellipse centre of ~1.57 times the local ellipse radius, suggesting a possible origin as a modified or partial outer basin ring. The ring structures of the Hellas and Argyre basins on Mars consist of an inwards-facing scarp at the outer ring, transitioning to a concave-upwards to sloping bench of lower topography and thinner crust, before reaching the main basin rim. Comparison of the azimuthally averaged topographic

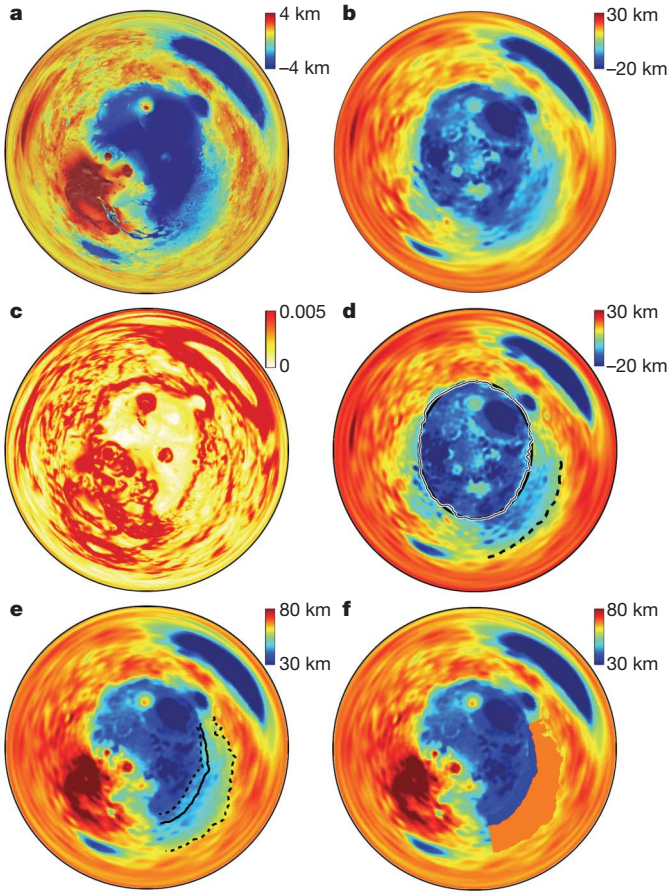


Figure 2 | Projected views of the Borealis basin. Projection in polar coordinates around the basin centre at 67° N, 208° E, showing the present-day topography and shaded relief of Mars (**a**), the modelled crustal root (**b**), and the topographic gradient at 4° wavelength (**c**). The traced dichotomy boundary is shown in **d** and compared with the best-fit ellipse (southern boundary of Arabia Terra denoted by dashed line). Outlines of the northern and southern edges of Arabia Terra (dotted lines; approximated using a threshold crustal thickness) are shown over a crustal thickness map[2] in **e**, along with the reconstructed basin rim required to restore the crustal thickness in Arabia Terra to the mean highlands value (solid line; see Methods). The reconstructed crustal thickness before basin modification in Arabia Terra is shown in **f**. Note that this projection preserves radial distance from the centre, and thus accurately represents the shape of the dichotomy boundary while distorting features not centred on the origin.

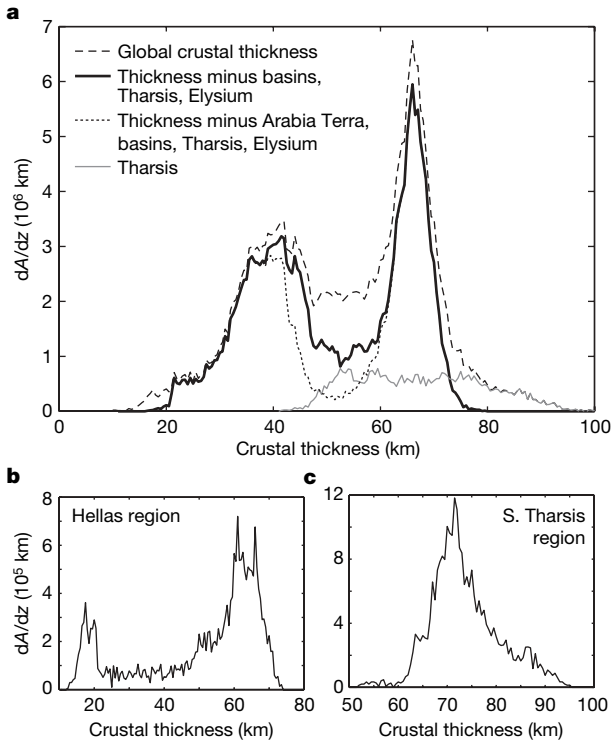


Figure 3 | Crustal thickness histograms. a, Global crustal thickness histogram (dashed), after removal of the major impact basins and volcanic rises (solid), and after removal of the anomalous Arabia Terra region as well (dotted). The histogram of the Tharsis region (excluding surrounding terrains) is shown in grey. For comparison, histograms are also shown of the Hellas impact basin and surrounding highlands (**b**), and the southern portion of Tharsis and the surrounding highlands (in order to avoid the competing effects of the superimposed dichotomy boundary beneath Tharsis; **c**). Histograms are presented as total area A per unit crustal thickness z, calculated in thickness increments of 0.5 km.

profile of the Borealis basin through Arabia Terra with profiles of Hellas and Argyre reveals a similarity in structure (Fig. 4). One model for the formation of multi-ring structures involves the inward flow of the ductile lower crust and mantle in the moments following the excavation of the transient cavity, leading to tectonic failure at the outer ring scarp[20]. A subtle change in the orientation of the dichotomy boundary at the eastern edge of Arabia Terra (Fig. 2c) suggests that this inward flow during ring formation may have displaced the basin rim northwards here. This lower crustal flow can also explain the northward shift of the dichotomy boundary at the Moho relative to the topographic boundary in Arabia Terra[1], and the associated gravity signature of the boundary[21]. This putative multi-ring structure does not fully encircle the basin, but a partial ring is also observed around the Caloris basin on Mercury[22]. The original basin rim location can be estimated by calculating the shortening required to restore the crust in Arabia Terra to the mean highlands thickness (Fig. 2e, f), leading to a corrected ring spacing of 1.42.

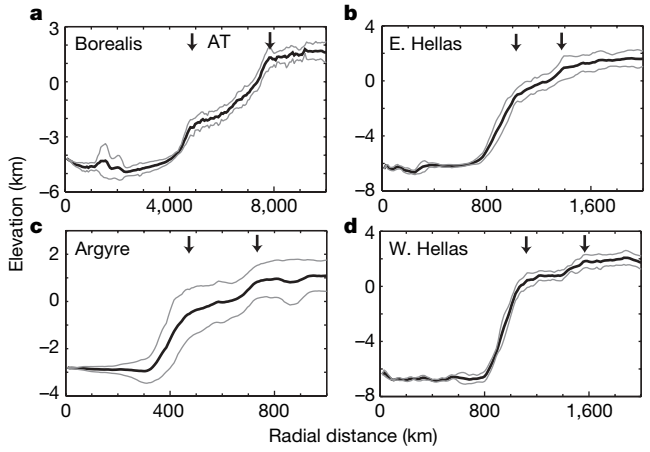


Figure 4 | Radial profiles of the Borealis (through Arabia Terra), Hellas and Argyre basins. Average profiles (black) and the 1σ variation (grey) were calculated from radial profiles at 1° increments. Borealis and Hellas were stretched in a basin-centred polar coordinate system to circularize the basins before averaging. Profiles of Hellas and Argyre avoided regions with obvious evidence of fluvial, volcanic, or subsequent impact modification (see Supplementary Fig. 6 for locations of profiles). The arrows indicate the approximate locations of the basin rim and outer ring.

One of the primary objections to the giant impact hypothesis is that the prodigious volume of melt produced by such an energetic impact would erase any signature of the basin[23]. However, recent simulations of oblique dichotomy-forming impacts[24] predict a smaller melt volume that is largely contained within the basin. Giant impact basins are thought to reach a near-isostatic state rapidly after the impact owing to the rebound of the basin floor, and the melt volume that can be contained within depends on the melt density (dominated by mantle material[23]) and the thickness of the surrounding crust. For the arbitrary case of a melt density intermediate between that of the unmelted crust and mantle, a crustal thickness of 70 km outside the basin would allow an isostatic magma pond up to 140 km in depth within the basin, corresponding to a global equivalent depth of 56 km of melt, compatible with the numerically predicted[24] global equivalent depth of 6–80 km. This local magma ocean would then differentiate to produce the present lowlands crust. Once formed, the crustal thickness signature of the dichotomy would not be expected to relax significantly[25], despite the high heat flow on early Mars[26]. Models suggest that the deepest basins are slowest to relax, on account of the resistance to flow of the thin basin-floor crust[27], explaining the lack of relaxation of Hellas and South Pole–Aitken.

By demonstrating that the northern lowlands are in fact elliptical in shape, we have removed the main obstacle to the impact hypothesis for the origin of the dichotomy. Recent numerical modelling work[24] has addressed the dynamical challenges to an impact origin. Although the alternative endogenic models remain viable, this work suggests that the giant impact hypothesis best explains the salient properties of the martian dichotomy. If the dichotomy was indeed formed by a giant impact, it would be the largest observed impact scar in the Solar System by a factor of ~4. Such large impacts were probably common during the waning stages of planetary accretion[28]. The formation of the Earth's Moon is attributed to a giant impact on the Earth by a Mars-sized body[29]. Similarly, the high density of Mercury may be the result of a giant impact that stripped the outer portions of its mantle[30]. These impacts were so catastrophic as to remove any trace of a basin from the surface. Although the details of giant impact basin formation and modification are poorly understood, it stands to reason that a basin might exist that spans the gap between the unambiguous Hellas and South Pole–Aitken basins, and the self-erasing Moon-forming and Mercury-stripping impacts.

METHODS SUMMARY

The gravity and topography of Mars were inverted using a spherical harmonic thin-shell model[14]. The gravity field[10] was expanded out to degree and order 60, with a cosine taper between degrees 55 and 60. The model assumes values of the mean crustal thickness (50 km), lithosphere thickness (50–200 km), crustal density (2,900 kg m^{-3}) and mantle density (3,500 kg m^{-3}) consistent with estimates for Mars, and values of Young's modulus (100 GPa) and Poisson's ratio (0.25) consistent with typical terrestrial values. The modelled bottom crustal thickness perturbations are assumed to be representative of the ancient isostatic crustal roots and thus to reflect the original crustal dichotomy before Tharsis loading. We then found the best-fit ellipse to the globally continuous dichotomy boundary by iterating the ellipse dimensions, centre coordinates, and orientation to minimize the root-mean-square misfit with the mapped boundary.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 20 December 2007; accepted 4 April 2008.

1. Zuber, M. T. *et al.* Internal structure and early thermal evolution of Mars from Mars Global Surveyor topography and gravity. *Science* **287,** 1788–1793 (2000).
2. Neumann, G. A., Lemoine, F. G., Smith, D. E. & Zuber, M. T. Marscrust3: A crustal thickness inversion from recent MRO gravity solutions. *Lunar Planet. Sci. Conf.* **39,** abstr. 2167 (2008).
3. Solomon, S. C. *et al.* New perspectives on ancient Mars. *Science* **307,** 1214–1220 (2005).
4. Frey, H. V., Roark, J. H., Shockey, K. M., Frey, E. L. & Sakimoto, S. E. H. Ancient lowlands on Mars. *Geophys. Res. Lett.* **29,** doi:10.1029/2001GL013832 (2002).
5. Wilhelms, D. E. & Squyres, S. W. The martian hemispheric dichotomy may be due to a giant impact. *Nature* **309,** 138–140 (1984).
6. Frey, H. & Shultz, R. A. Large impact basins and the mega-impact origin for the crustal dichotomy on Mars. *Geophys. Res. Lett.* **15,** 229–232 (1988).
7. Zhong, S. & Zuber, M. T. Degree-1 mantle convection and the crustal dichotomy on Mars. *Earth Planet. Sci. Lett.* **189,** 75–84 (2001).
8. Roberts, J. H. & Zhong, S. Degree-1 convection in the martian mantle and the origin of the hemispheric dichotomy. *J. Geophys. Res.* **111,** E06013, doi:10.1029/2005JE002668 (2006).
9. Elkins-Tanton, L. T., Hess, P. C. & Parmentier, E. M. Possible formation of ancient crust on Mars through magma ocean processes. *J. Geophys. Res.* **110,** E12S01, doi:10.1029/2005JE002480 (2005).
10. Konopliv, A. S. *et al.* MROMGM23C gravity model. *NASA Planet. Data Sys.* ⟨http://pds.jpl.nasa.gov⟩ (submitted).
11. Smith, D. E. *et al.* Mars Orbiter Laser Altimeter: Experiment summary after the first year of global mapping of Mars. *J. Geophys. Res.* **106** (E10), 23689–23722 (2001).
12. McGill, G. E. & Squyres, S. W. Origin of the martian crustal dichotomy: Evaluating hypotheses. *Icarus* **93,** 386–393 (1991).
13. Phillips, R. J. *et al.* Ancient geodynamics and global-scale hydrology on Mars. *Science* **291,** 2587–2591 (2001).
14. Banerdt, W. B. & Golombek, M. P. Tectonics of the Tharsis region of Mars: Insights from MGS topography and gravity. *Lunar Planet. Sci. Conf.* **31,** abstr. 2038 (2000).
15. Andrews-Hanna, J. C., Zuber, M. T. & Hauck, S. A. Strike-slip faults on Mars: Observations and implications for global tectonics and geodynamics. *J. Geophys. Res.* doi:10.1029/2008JE002980 (in the press).
16. Arkami-Hamed, J. A coherent model of the crustal magnetic field of Mars. *J. Geophys. Res.* **109,** E09005, doi:10.1029/2004JE002265 (2004).
17. Smrekar, S. E., McGill, G. E., Raymond, A. & Dimitriou, A. M. Geologic evolution of the martian dichotomy in the Ismenius area of Mars and implications for plains magnetization. *J. Geophys. Res.* **109,** E11002, doi:10.1029/2004JE002260 (2004).
18. Harder, H. & Christensen, U. A one-plume model of martian mantle convection. *Nature* **380,** 507–509 (1996).
19. Melosh, H. J. *Impact Cratering: A Geologic Process* (Oxford Univ. Press, New York, 1989).
20. Melosh, H. J. & McKinnon, W. B. The mechanics of ringed basin formation. *Geophys. Res. Lett.* **5,** 985–988 (1978).
21. Kiefer, W. S. Gravity, topography, and tectonic segmentation of the martian hemispheric dichotomy: Evidence for multiple formation mechanisms. *Lunar Planet. Sci. Conf.* **38,** abstr. 1470 (2007).
22. Strom, R. *et al.* Tectonism and volcanism on Mercury. *J. Geophys. Res.* **80,** 2478–2507 (1975).
23. Tonks, W. B. & Melosh, H. J. Magma ocean formation due to giant impacts. *J. Geophys. Res.* **98,** 5319–5333 (1993).
24. Marinova, M. M., Aharonson, O. & Asphaug, E. Mega-impact formation of the Mars hemispheric dichotomy. *Nature* doi:10.1038/nature07070 (this issue).
25. Nimmo, F. & Stevenson, D. J. Estimates of martian crustal thickness from viscous relaxation of topography. *J. Geophys. Res.* **106,** 5085–5098 (2001).
26. Parmentier, E. M. & Zuber, M. T. Early evolution of Mars with mantle compositional stratification or hydrothermal crustal cooling. *J. Geophys. Res.* **112,** E02007, doi:10.1029/2005JE002626 (2007).
27. Mohit, P. S. & Phillips, R. J. Viscous relaxation on early Mars: A study of ancient impact basins. *Geophys. Res. Lett.* **34,** L21204, doi:10:1029/2007GL031252 (2007).
28. Wetherill, G. W. Occurrence of giant impacts during the growth of the terrestrial planets. *Science* **228,** 877–879 (1985).
29. Canup, R. M. & Esposito, L. W. Accretion of the Moon from an impact-generated disk. *Icarus* **119,** 427–446 (1996).
30. Wetherill, G. W. in *Mercury* (eds Vilas, F., Chapman, C. R. & Mathews, M. S.) 670–691 (Univ. Arizona Press, Tucson, 1988).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Melosh for a review. This work was supported by grants to M.T.Z. from the Mars Reconnaissance Orbiter project, operated under the auspices of the NASA Mars Program.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.C.A.-H. (jhanna@mit.edu).

METHODS

Gravity–topography inversion. The observed martian gravity and topography were inverted under the assumption of crustal thickness-compensated flexure, using the model of Banerdt[31] (see refs 14 and 31 for full details). By taking into account the strength of the lithosphere, the model solves for the crustal thickness and subdivides the crust into that portion which is supported by the membrane and flexural stresses in the lithosphere and that which is in isostatic balance. The model results are represented as a crust of uniform thickness, to which thickness perturbations are added at the top and bottom boundaries (the loads and isostatic roots, respectively) that, together with the resulting displacements, reproduce the observed gravity and topography. Note that the 'isostatic root' as defined here removes the flexural contribution from the conventional crustal root, defined as that portion of the crust below a particular equipotential. The pre-Tharsis crustal dichotomy was probably isostatically compensated, with a buoyant crustal root supporting the topography of the highlands, resulting in zero flexure. In contrast, Tharsis would have been largely a flexurally supported load (Supplementary Fig. 1).

The primary uncertainty is the choice of lithosphere thickness. Although admittance studies of the lithosphere thickness supporting shorter-wavelength loads suggest a thin lithosphere ($<$50 km) during the Noachian[32,33], models of Tharsis loading suggest that a lithosphere thickness of $\sim$100 km is required to reproduce the observed gravity anomalies[13] and explain the tectonic history of the rise[14,15]. As a possible explanation for this discrepancy, we suggest that early Mars may have had an upper crustal lithosphere separated from a mantle lithosphere by a ductile lower crust[34,35]. Short-wavelength, low-amplitude deformation beneath smaller scale loads may be supported entirely by the stresses within the elastic upper crust, with the flexure easily accommodated by the thin, ductile lower crust. In contrast, the long-wavelength, large-amplitude Tharsis deformation may involve support from both the mantle and crustal lithosphere. We consider lithosphere thickness between 50 and 200 km.

The isostatic roots beneath Tharsis are largely negative for the 50-km lithosphere (Supplementary Fig. 2a), thus predicting a topographic depression in the Tharsis region before loading. The fact that these negative crustal roots mirror the present-day topography of Tharsis suggests that this is an artefact of assuming an overly thin lithosphere, rather than representing the pre-Tharsis crustal structure. For lithosphere thicknesses ranging between 100 and 200 km, the pre-Tharsis dichotomy boundary is clearly expressed and lies in roughly the same location (Supplementary Fig. 2b–d). Because the latitude of the sub-Tharsis dichotomy varies little across the rise, we also consider longitudinally averaged profiles of the isostatic root through the Tharsis region (Supplementary Fig. 3). The maximum slope delineating the dichotomy boundary scarp occurs in the same location for all lithosphere thicknesses, demonstrating that the inferred location of the sub-Tharsis dichotomy boundary is generally insensitive to this choice.

Boundary mapping and ellipse fit. We mapped the dichotomy boundary in a cylindrical projection. The location of the dichotomy boundary was traced outside of Tharsis using the topography and topographic gradient, and beneath Tharsis using the isostatic root and its gradient (Supplementary Fig. 4). The gradient of the root shows the dichotomy boundary clearly continuing from the region west of Tharsis across the rise. Interestingly, the boundary scarp is more distinct beneath the western portion of Tharsis than in Arabia Terra. The dichotomy boundary north of Arabia Terra as seen in the crustal root is less distinct and seems to lie somewhat further north than indicated by the topography[1]. This could be a result of lower crustal flow during the suggested formation of Arabia Terra as a basin ring structure during the impact, owing to imperfect coupling between the lower ductile and upper brittle crust. Alternatively, it could result from relaxation of the crustal root later in Mars history, or from weathering and retreat of the topographic dichotomy boundary. We trace the more distinct topographic dichotomy boundary in this region. The uncertainties in the boundary location are comparable to the 350-km resolution of the gravity data beneath Tharsis, and less than 100 km outside Tharsis.

After global mapping of the dichotomy boundary, we found the best-fit ellipse by iteratively adjusting the location of the ellipse centre, major and minor axes, and orientation of the major axis to minimize the root-mean-square misfit (Supplementary Fig. 5). Tracing the boundary north of Arabia Terra using the isostatic root rather than topography results in a best-fitting ellipse with major and minor axes of 10,600 and 8,100 km, respectively. Alternatively, the reconstructed rim in Arabia Terra leads to ellipse axes of 10,900 and 8,800 km.

Basin profile measurement. In comparing the radial profiles of the Borealis basin with profiles of Argyre and Hellas, it was necessary to circularize the basins in the basin-centred coordinate system. We then took azimuthally averaged profiles across the dichotomy in the Arabia Terra region (Supplementary Fig. 6). The rim of Hellas has been modified in places by fluvial and volcanic processes, so profiles were averaged in two sectors of the rim that seem to have been largely preserved. The rim of Argyre is superimposed by one prominent impact crater, and this was similarly excluded from the profiles.

Arabia Terra rim reconstruction. We calculated the reconstructed basin rim location in Arabia Terra by assuming that the crustal thickness in Arabia Terra was originally equal to that in the southern highlands, before modification during basin ring formation. Upper and lower crustal thickness thresholds were defined to bracket Arabia Terra (40 and 60 km, respectively). The crustal thickness within Arabia Terra was then integrated in wedges radiating out from the basin centre, and the shortening required to increase the crustal thickness to the mean highlands value was calculated.

31. Banerdt, W. B. Support of long-wavelength loads on Venus and implications for internal structure. *J. Geophys. Res.* **91,** 403–419 (1986).
32. McGovern, P. J. *et al.* Localized gravity/topography admittance and correlation spectra on Mars: implications for regional and global evolution. *J. Geophys. Res.* **107,** 5136, doi:10.1029/2002JE001854 (2002).
33. McGovern, P. J. *et al.* Correction to ''Localized gravity/topography admittances and correlation spectra on Mars: Implications for regional and global evolution''. *J. Geophys. Res.* **109,** doi:10.1029/2004JE002286 (2004).
34. Grott, M. & Breuer, D. The evolution of the martian elastic lithosphere and implications for crustal and mantle rheology. *Icarus* **193,** 503–515 (2008).
35. Montessi, L. G. J. & Zuber, M. T. Clues to the lithospheric structure of Mars from wrinkle ridge sets and localization instability. *J. Geophys. Res.* **108,** 5048, doi:10.1029/2002JE001974 (2003).

LETTERS

Mega-impact formation of the Mars hemispheric dichotomy

Margarita M. Marinova[1], Oded Aharonson[1] & Erik Asphaug[2]

The Mars hemispheric dichotomy is expressed as a dramatic difference in elevation, crustal thickness and crater density between the southern highlands and northern lowlands (which cover ~42% of the surface)[1,2]. Despite the prominence of the dichotomy, its origin has remained enigmatic and models for its formation largely untested[3–5]. Endogenic degree-1 convection models with north–south asymmetry are incomplete in that they are restricted to simulating only mantle dynamics and they neglect crustal evolution, whereas exogenic multiple impact events are statistically unlikely to concentrate in one hemisphere[6]. A single mega-impact of the requisite size has not previously been modelled. However, it has been hypothesized that such an event could obliterate the evidence of its occurrence by completely covering the surface with melt[7] or catastrophically disrupting the planet[3,8]. Here we present a set of single-impact initial conditions by which a large impactor can produce features consistent with the observed dichotomy's crustal structure and persistence. Using three-dimensional hydrodynamic simulations, large variations are predicted in post-impact states depending on impact energy, velocity and, importantly, impact angle, with trends more pronounced or unseen in commonly studied smaller impacts[9]. For impact energies of ~(3–6) × 10^{29} J, at low impact velocities (6–10 km s^{-1}) and oblique impact angles (30–60°), the resulting crustal removal boundary is similar in size and ellipticity to the observed characteristics of the lowlands basin. Under these conditions, the melt distribution is largely contained within the area of impact and thus does not erase the evidence of the impact's occurrence. The antiquity of the dichotomy[10] is consistent with the contemporaneous presence of impactors of diameter 1,600–2,700 km in Mars-crossing orbits[3], and the impact angle is consistent with the expected distribution[11].

The martian dichotomy may be defined by topographical, morphological and structural characteristics. Isostatic modelling combining gravity and topography have provided a description of global crustal thickness in which the northern lowlands are distinguished from the southern highlands by a reduction in crustal thickness of ~30 km (ref. 1). By accounting for lithospheric stresses, it is possible to compute the effects of overlying loads, in particular of the largest load represented by the Tharsis province. When the loads are separated, the lowlands are remarkably well described by an ellipse with dimensions ~10,650 km × ~8,520 km (ellipticity ~1.25) (ref. 2). The boundary is expressed as steep scarps in some longitudes and as gentle slopes in others[3,12,13]; significant crustal thickening is not observed at the boundary. Geochemical evidence and surface-crater densities show that the dichotomy formed within the first 50 Myr of Solar System formation, with little mantle-crust remixing since[1,10,14]. Subsequent events, such as known impact-basin formation, have modified the dichotomy boundary.

The mega-impact formation hypothesis is supported by geologic evidence including massifs and narrow plateaux concentric to the dichotomy boundary[3], steep scarps at the boundary, and by the similarity of the lowlands to other large impact basins such as South Pole–Aitken basin on the Moon, Caloris basin on Mercury and Hellas basin on Mars. The impact hypothesis has previously been challenged by several arguments. First, by the expectation that at the relevant energy, the impact would disrupt the planet sufficiently to effectively erase evidence of the event[3]. Second, by the circularity of craters for all but the most oblique angles for smaller impacts[15]. Third, by the lack of crustal thickening in an annulus around the basin, typical for smaller impacts. However, craters resulting from planetary-scale impacts have until now not been accurately modelled. This class of impacts is distinguished from the more thoroughly studied smaller impacts, which effectively form in a half-space target, in part because of the importance of surface curvature in the larger size regime and the larger fractional size of the projectile relative to the target.

Single, planetary-scale impact events are simulated using a three-dimensional self-gravitating smoothed particle hydrodynamics (SPH) code[16–18]. Our simulations sample a large parameter space, with impact energies of (0.1–5.9) × 10^{29} J, which is representative of, according to traditional scaling laws[3,9], nominal impact crater diameters of 4,000–12,000 km. For comparison, the energy of the Moon-forming impact[18] was ~10^{31} J. For each impact energy, we consider impact velocities of 6–50 km s^{-1}, ranging from near escape velocity to twice Mars's orbital velocity, and impact angles of 0 (head-on), 15, 30, 45, 60 and 75° for each velocity (Supplementary Information). For this parameter space, impactor diameters range from 400 to 2,700 km. Figure 1 schematically shows a summary of the results and the 'sweet spot' simulations that produce a crustal excavation feature remarkably similar to the lowlands.

The pre-impact resolution (particle size or smoothing length) is 118 km for $N = 200{,}000$ particles. The model uses the semi-empirical Tillotson equation of state[19] (EOS). We derived EOS parameters to approximate the behaviour of olivine, to match the planet's pressure–density profile. The olivine EOS results in a realistic early Mars internal energy–pressure profile, allowing calculation of post-impact melt using the pressure-dependent forsterite liquidus curve as an internal energy melting threshold[20]. The pre-impact planet has no initial spin: Mars's current rotational period is long compared with the timescale of the impact process. The crust is defined as the planet's pre-impact outermost particle layer, resulting in a crustal thickness of ~140 km, compared with recent estimates of 5–90 km (ref. 1). Because of the large particle size, the simulations cannot directly resolve the crustal thickness. However, the region of complete crustal removal may be mapped and the boundary of the crustal anomaly is expressed over a lateral distance of only several resolution elements.

[1]California Institute of Technology, Division of Geological and Planetary Sciences, MC 150-21, Pasadena, California 91125, USA. [2]Earth Sciences Department, University of California, Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA.

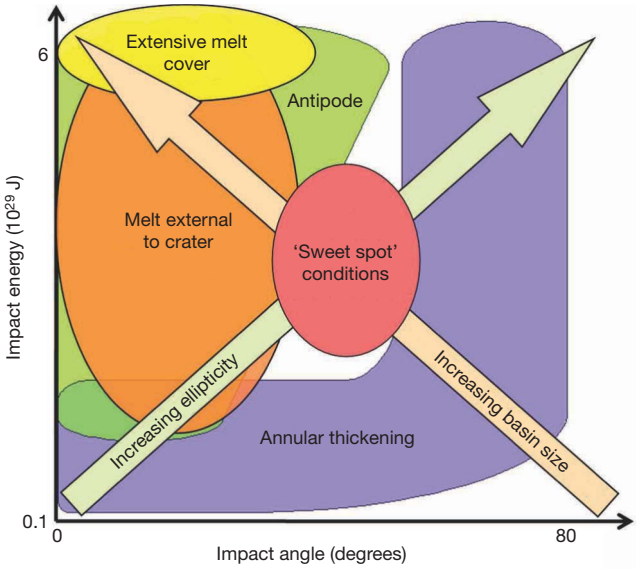


Figure 1 | Summary of simulation results. Shown are the impact characteristics resulting in extensive surface melt cover (>25% of the surface), significant melt outside the crustal excavation boundary, presence of antipodal crustal disruption, presence of a thickened annulus of crust around the crustal excavation boundary, and the directions of increase in ellipticity and basin size. The results at a given energy are averaged over impact velocity. A 'sweet spot' of impact conditions emerges for which the resulting simulation characteristics closely match the observed Mars dichotomy features[2]. A compatible hypothesis is found at an impact energy of $\sim 3 \times 10^{29}$ J, velocity ~ 6 km s^{-1} and, importantly, an impact angle of $\sim 45^{\circ}$. These parameters represent probable impact conditions in the early Solar System[3,11].

Thus the computed crustal excavation boundary size is a robust result. In addition to this boundary, we consider the integrated amount and spatial distribution of melt, crustal thickening and the extent of antipodal disruption.

The distribution of crust and surface melt are calculated as a fraction of the material within the top 150 km. An ellipse is fitted to the crustal excavation boundary (the contour of 50% crustal fraction) in polar coordinates, with the origin centred on the excavated region. Our analysis of the impact melt and its distribution shows that previous assumptions about melting during planetary-scale cratering events have been oversimplified.

In contrast to smaller, half-space craters, whose size and melt production dominantly scale with the impact energy[21], for planetary-scale impacts we find that impact velocity and impact angle fundamentally affect the crustal excavation boundary, its ellipticity, and the amount and distribution of melt. In particular, we identify possible impacts that are consistent with the crustal distribution of Mars.

Planetary-scale impacts penetrate into the mantle. The resulting rarefaction wave completely removes the surrounding crust, which re-impacts elsewhere on the planet or is ejected to space. The size of the crustal excavation boundary is representative of the size of the crustal thickness dichotomy that is likely to remain, neglecting later geologic crater modification. Simulation results show that the crustal excavation boundary size increases with increasing impact energy. For a given impact energy, the boundary size decreases with increasing velocity and with increasingly oblique impacts (Figs 2 and 3a). For smaller, half-space impact craters, a deviation in circularity is only present for highly oblique impacts[9,15] (>80°). In contrast, our planetary-scale impact simulations show that with increasing impact energy, the removed crustal region becomes significantly elongated at relatively shallow angles (Fig. 3b).

The pattern of crustal redistribution depends upon impact angle. Although angles above ~60° result in a distinct rim-like feature, less oblique impacts (<45°) produce widespread crustal thickening but no short length-scale variations, in agreement with dichotomy characteristics (Fig. 4; contrast with Supplementary Information). In cases with high ejection velocity, the flight path of ejected material is of the order of the radius of the planet; thus the ejected material is distributed globally.

Melt production and distribution are also strongly dependent on impact energy, velocity and angle. The total amount of melt increases with increasing impact energy, and at constant energy and low impact angles exhibits a weak maximum at intermediate velocities (10–20 km s^{-1}). Melt significantly decreases with increasing impact angle. As an example, for a nominal 10,000-km crater, head-on (0°) impacts produce a Mars global equivalent layer (GEL) melt depth

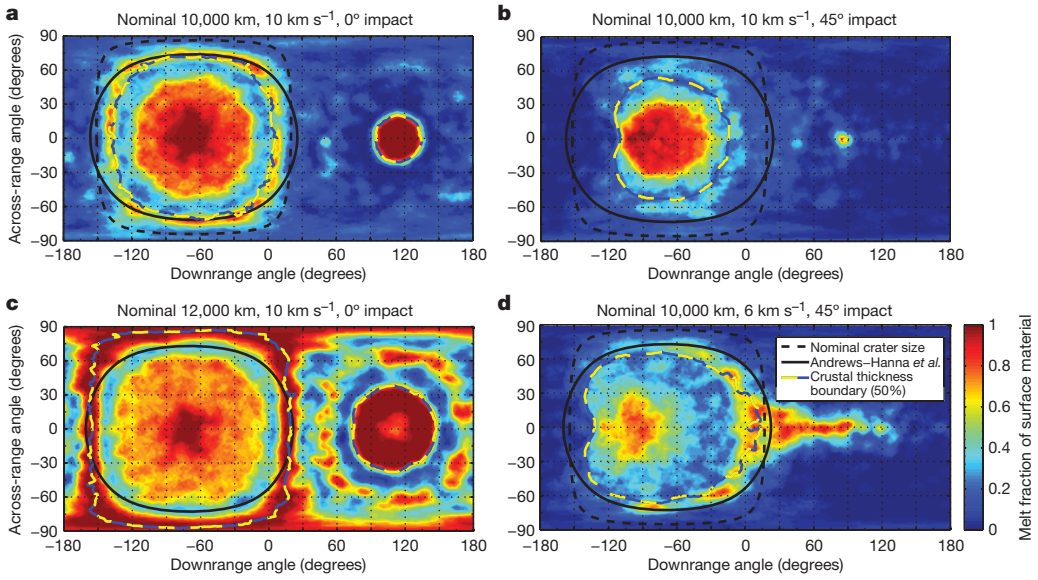


Figure 2 | Change in melt distribution and crustal removal boundary with impact characteristics. Crustal excavation boundary, nominal crater size and a fit by Andrews-Hanna *et al.*[2] to the dichotomy boundary are overlaid. The melt distribution is computed at a 2° resolution and smoothed over a 10° diameter cap area. The surface melt cover fractions are 25%, 8%, 71% and 12%, respectively. Note the changes in features with impact energy (nominal crater size), velocity and angle. The planet has been rotated to centre the excavation boundary at approximately −60° downrange angle.

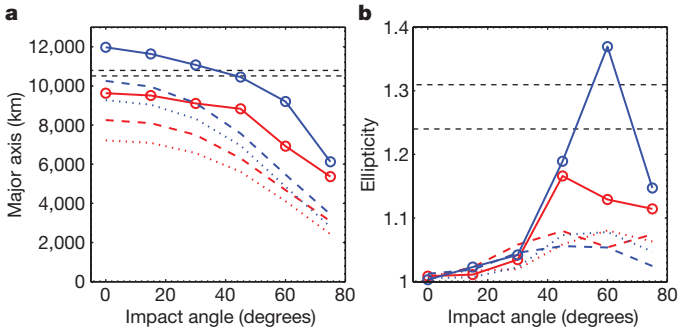


Figure 3 | Major axis and ellipticity for impact energies of 3. 1×10^{29} J and 5.9×10^{29} J (red and blue, respectively). a, Excavated cavity major axis; **b**, ellipticity. Shown are impact velocities of 6 km s^{-1} (solid line), 10 km s^{-1} (dashed line) and 50 km s^{-1} (dotted line). Major axes and ellipticities of mapped dichotomy boundary ellipse fits[2] are shown (black dashed lines), representing the range of possible boundary locations (reported uncertainty of ± 100 km). A 'sweet spot' emerges for these impact energies and at impact velocities of 6–10 km s^{-1} and impact angles of 30–60°.

of 60–80 km (depending on impact velocity), whereas 75° impacts produce a GEL melt depth of only 6–20 km. The vaporized mass is less than 1% of the molten mass.

Global melt depths of tens of kilometres have been argued to be sufficient to erase the signature of the dichotomy[7]; however, GEL depths do not represent the highly heterogeneous distribution of melt. The distribution varies with impact characteristics. For all but the highest energies (nominal crater size ≤10,000 km), melt is largely contained within the crustal excavation boundary and extends to depth (Figs 1 and 2). Depending on impact angle, 50–70% of the melt resides inside the excavation boundary, 25–30% is deposited outside the boundary and the remainder is ejected from the planet. Most re-deposited material is of crustal composition and results in a thickening of up to ~60% compared with the original crustal thickness.

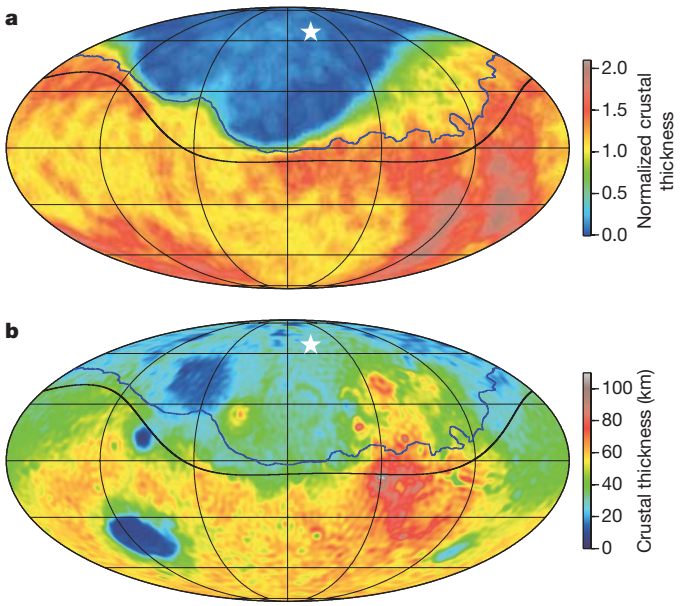


Figure 4 | A favoured impact hypothesis compared with Mars's crustal thickness. Post- to pre-impact simulation crustal thickness ratio (**a**), and model thicknesses (based on gravity and topography[10], revised by Neumann *et al.*, manuscript in preparation) (**b**). Superimposed are the Andrews-Hanna *et al.* dichotomy boundary[2] (black line) and the crustal excavation boundary from the simulation results (blue line). Impact simulation characteristics: 3.1×10^{29} J (nominal 10,000-km crater), 6 km s^{-1}, 45°, impactor diameter 2,230 km. Crustal excavation boundary centre[2] (star) shown at 66° N, 206° E. In **a**, the crustal thickness is computed at a 2° resolution and smoothed over a 10°-diameter cap area.

In areas where crust is removed and the mantle melts, fresh crust that crystallizes is likely to leave a difference in crustal thickness. The amount of mantle melt, and hence the thickness of the new crustal layer, is dependent on impact conditions.

For highly energetic and fast impacts, the shock wave produced is sufficiently strong to induce antipodal effects including crustal removal and melting. These are inconsistent with the lack of observed topographic, gravitational or magnetic anomalies antipodal to the proposed impact location. Thus we only consider viable simulations that produce antipodal features smaller than 10° in diameter.

We consider the effect of numerical resolution on the simulation results. The resolution and fidelity of post-impact crustal features in these simulations is higher than that of previous three-dimensional SPH studies. For simulations with a particle smoothing length of 150 km, the basin major axis, ellipticity, antipode size and melt cover differ from the nominal 118-km resolution simulations by an average of −8%, −1%, −28% and −16%, respectively, for nominal 10,000- and 12,000-km craters. Thus the qualitative conclusions are robust.

Combining the crust and melt distribution results, we find a 'sweet spot' in parameter space, where the simulations show striking similarity to the observed Mars dichotomy features (Figs 1, 3 and 4). Importantly, this range represents impact conditions that are probable in light of the age of the dichotomy[10] and probability distribution of the impact angle[11,22]. This parameter space 'sweet spot' is at impact energies of ~(3–6) × 10^{29} J, impact angles of 30–60° and impact velocities of 6–10 km s^{-1}, which imply impactor diameters of 1,600–2,700 km. These favoured simulation conditions encompass the range of uncertainty in the geometry of the observed crustal anomaly. The early age of the dichotomy is consistent with the expected timing of the influx of large impactors. These objects are also expected to have similar orbital velocities[23], resulting in impacts at or slightly above Mars's escape velocity (≥5 km s^{-1}). The most likely impact angle[11] is 45°.

Results from the large parameter space explored by the simulations provide new insights pertinent to global-scale impact processes thought to prevail in the early Solar System. Our simulations provide quantitative constraints for the previously only hypothesized extent of surface melting, planetary disruption and crustal removal as a function of impact energy and geometric characteristics. The predicted melt distribution over the surface may provide a heterogeneous geochemical signature observable by future Mars missions.

METHODS SUMMARY

SPH is a lagrangian method in which matter is represented by point masses smoothed over a particle radius (smoothing length), with density and internal energy computed according to kernel-weighted summation and by the conservation of mass, momentum and energy[16]. Pressure, as a function of internal energy and density, is computed with the Tillotson EOS, and pressure gradients and self-gravitating forces accelerate the particles. Our simulations conserve energy and angular momentum to better than 1 part in 2,000. Simulations are run for 26 h of model time, after which the r.m.s. particle velocity does not appreciably oscillate. We assume an olivine composition of $Fo_{75}Fa_{25}$ (ref. 24). Density ($\rho_0 = 3{,}500$ kg m^{-3}) (ref. 25), bulk modulus ($K = 131$ GPa) (ref. 26), heat capacity[26] and heat of vaporization ($H_{vap} = 10.013$ MJ kg^{-1}) (ref. 27) are measured material values; the nonlinear Tillotson compressive term (B) and two of the Tillotson EOS fitting parameters (b, U_0) are set to the average of those published for basalt, granite, anorthosite low- and high-pressure phases, and andesite ($B = 49$ GPa, $b = 1.4$, $U_0 = 550$ MJ kg^{-1}); b varies by only 8%. The remaining Tillotson EOS fitting parameters are identical for all given rocky materials ($a = 0.5$, $\alpha = 5$, $\beta = 5$). The olivine Hugoniot internal energy curve is on average 15% lower and 11% higher than the experimentally determined pure forsterite and fayalite curves, respectively, for 0–200 GPa. Using a forsterite EOS with Tillotson parameters fitted to the experimental curve results, on average, in 8% more melt (impacts of 8,000–12,000 km) and similar melt distribution. Both the mantle and crust are composed of olivine because a single-particle basalt layer would be numerically unresolved. The core is composed of iron and the impactor of basalt. The SPH code was modified to initialize with randomly distributed particles of prescribed composition, internal energy, pressure and mass as a function of radial position. Transient oscillations are damped

during a relaxation period run. The initial internal energy–pressure profile is set to that of hydrostatic equilibrium, whereas the surface and core-mantle boundary temperatures are set to those of parameterized convection models of Mars[28]. The internal energy–pressure–density profile is computed assuming adiabatic compression into the planet (core radius 1,600 km, central pressure 50 GPa, compatible with models[29,30]). The crustal excavation boundary size is a robust result: for a nominal 10,000-km crater, fitting the 20% and 80% crustal-fraction contours changes the boundary size by −9% and 12%, respectively.

Received 5 December 2007; accepted 23 April 2008.

1. Zuber, M. T. The crust and mantle of Mars. *Nature* **412,** 220–227 (2001).
2. Andrews-Hanna, J. C., Zuber, M. T. & Banerdt, W. B. The Borealis basin and the origin of the martian crustal dichotomy. *Nature* doi:10.1038/nature07011 (this issue).
3. Wilhelms, D. E. & Squyres, S. W. The martian hemispheric dichotomy may be due to a giant impact. *Nature* **309,** 138–140 (1984).
4. Zhong, S. J. & Zuber, M. T. Degree-1 mantle convection and the crustal dichotomy on Mars. *Earth Planet. Sci. Lett.* **189,** 75–84 (2001).
5. Frey, H. & Schultz, R. A. Large impact basins and the mega-impact origin for the crustal dichotomy on Mars. *Geophys. Res. Lett.* **15,** 229–232 (1988).
6. McGill, G. E. & Squyres, S. W. Origin of the martian crustal dichotomy – evaluating hypotheses. *Icarus* **93,** 386–393 (1991).
7. Hart, S. H., Nimmo, F., Korycansky, D. & Agnor, C. Probing the giant impact hypothesis of the martian crustal dichotomy. *Proc. 7th Int. Conf. Mars* abstr. 3332 (2007).
8. Nimmo, F. & Tanaka, K. Early crustal evolution of Mars. *Annu. Rev. Earth Planet. Sci.* **33,** 133–161 (2005).
9. Melosh, H. J. *Impact Cratering: A Geologic Process* (Oxford Univ. Press, New York, 1989).
10. Solomon, S. C. *et al.* New perspectives on ancient Mars. *Science* **307,** 1214–1220 (2005).
11. Shoemaker, E. M. in *Physics and Astronomy of the Moon* (ed. Kopal, Z.) 283–359 (Academic Press, New York, 1962).
12. Smith, D. E. *et al.* The global topography of Mars and implications for surface evolution. *Science* **284,** 1495–1503 (1999).
13. Aharonson, O., Zuber, M. T. & Rothman, D. H. Statistics of Mars' topography from the Mars Orbiter Laser Altimeter: slopes, correlations, and physical models. *J. Geophys. Res.* **106** (E10), 23723–23735 (2001).
14. Frey, H. V. *et al.* Ancient lowlands on Mars. *Geophys. Res. Lett.* **29,** doi:10.1029/2001GL013832 (2002).
15. Gault, D. E. & Wedekind, J. A. Experimental impact craters formed in water – gravity scaling realized. *Trans. Am. Geophys. Union* **59,** 1121 (1978).
16. Benz, W. in *Proc. NATO Adv. Res. Worksh. Numer. Modell. Nonlin. Stellar Puls.* (ed. Buchler, J. R.) 1–54 (Kluwer Academic, Boston, 1990).
17. Benz, W., Slattery, W. L. & Cameron, A. G. W. The origin of the Moon and the single-impact hypothesis. 1. *Icarus* **66,** 515–535 (1986).
18. Canup, R. M. & Asphaug, E. Origin of the Moon in a giant impact near the end of the Earth's formation. *Nature* **412,** 708–712 (2001).
19. Tillotson, J. H. *Metallic Equations of State for Hypervelocity Impact*. Report No. GA-3216, July 18 (General Atomic, San Diego, California, 1962).
20. Asimow, P. D. Magmatism and the evolution of the Earth's interior, in *Goldschmidt Conference Abstracts*, A40 ⟨http://www.goldschmidt2007.org⟩ (2007).
21. Cintala, M. J. & Grieve, R. A. Scaling impact melting and crater dimensions: implications for the lunar cratering record. *Meteorit. Planet. Sci.* **33,** 889–912 (1998).
22. Gilbert, G. K. The Moon's face, a study of the origin of its features. *Bull. Phil. Soc. Wash.* **12,** 241–292 (1893).
23. Canup, R. M. & Agnor, C. B. in *Origin of the Earth and Moon* (eds. Canup, R. M. & Righter, K.) 113–129 (Univ. Arizona Press, Tuscon, Arizona, 2000).
24. Sanloup, C., Jambon, A. & Gillet, P. A simple chondritic model of Mars. *Phys. Earth Planet. Inter.* **112,** 43–54 (1999).
25. Klein, C. *The Manual of Mineral Science* 22nd edn, 491–495 (Wiley, New York, 2002).
26. Anderson, D. L. & Isaak, D. G. in *Mineral Physics and Crystallography: A Handbook of Physical Constants* (ed. Ahrens, T. J.) 64–97 (American Geophysical Union, Washington DC, 1995).
27. Hashimoto, A. Evaporation metamorphism in the early solar nebula – evaporation experiments on the melt FeO-MgO-SiO_2-CaO-Al_2O_3 and chemical fractionations of primitive materials. *Geochem. J.* **17,** 111–145 (1983).
28. Hauck, S. A. & Phillips, R. J. Thermal and crustal evolution of Mars. *J. Geophys. Res.* **107** (E7), doi:10.1029/2001JE001801 (2002).
29. Yoder, C. F. *et al.* Fluid core size of Mars from detection of the solar tide. *Science* **300,** 299–303 (2003).
30. Bertka, C. M. & Fei, Y. W. Density profile of an SNC model martian interior and the moment-of-inertia factor of Mars. *Earth Planet. Sci. Lett.* **157,** 79–88 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank F. Nimmo, M. Zuber, J. Andrews-Hanna and R. Canup for discussions, J. Melosh for comments, and S. Squyres for suggesting the problem and the approach more than a decade ago. This work was supported by the Henshaw Fellowship, the Natural Sciences and Engineering Research Council of Canada and the Canadian Space Agency.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.M.M. (mmm@caltech.edu).

LETTERS

Implications of an impact origin for the martian hemispheric dichotomy

F. Nimmo[1], S. D. Hart[1], D. G. Korycansky[1] & C. B. Agnor[2]

The observation that one hemisphere of Mars is lower and has a thinner crust than the other (the 'martian hemispheric dichotomy')[1–3] has been a puzzle for 30 years. The dichotomy may have arisen as a result of internal mechanisms such as convection[4,5]. Alternatively, it may have been caused by one[6] or several[7] giant impacts, but quantitative tests of the impact hypothesis have not been published. Here we use a high-resolution, two-dimensional, axially symmetric hydrocode[8,9] to model vertical impacts over a range of parameters appropriate to early Mars. We propose that the impact model, in addition to excavating a crustal cavity of the correct size, explains two other observations. First, crustal disruption[10] at the impact antipode is probably responsible for the observed antipodal decline in magnetic field strength[11]. Second, the impact-generated melt forming the northern lowlands crust is predicted to derive from a deep, depleted mantle source. This prediction is consistent with characteristics of martian shergottite meteorites[12,13] and suggests a dichotomy formation time ~100 Myr after martian accretion[13], comparable to that of the Moon-forming impact on Earth[14].

It has been proposed[6], mainly on the basis of geological observations, that the northern lowlands of Mars were generated by a single giant impact with a kinetic energy of $\sim 10^{29}$ J. The hypothesized impact occurred at roughly 170° E, 50° N and resulted in the partial or total removal of pre-existing crust from a basin 3,850 km in radius. Later spacecraft measurements[1,2] confirmed a relatively abrupt increase in both elevation and crustal thickness 50–80° (3,000–4,750 km) from the putative impact centre, and an antipodal decline in magnetic field strength (Fig. 1). Also, cratering statistics suggest that the northern lowlands were formed early, at around 4.4 Gyr BP[3,15].

We have simulated giant impacts on Mars using the eulerian numerical hydrocode Zeus[8], modified to model impact events[9]. We verified that the code reproduced previous results for transient impact crater behaviour and also benchmarked it against lagrangian smoothed-particle hydrodynamics (SPH) results (see Supplementary Information). Individual impactor and simulated Mars characteristics are given in Supplementary Table 1 and Fig. 2, respectively. The excavation and redistribution of crustal material was tracked using passive tracer particles with 5-km vertical spacing in the top 50 km of the crust.

Because the martian crust has an average thickness of roughly 50 km (ref. 2), high spatial resolution is required to investigate crustal excavation and redistribution. We therefore focused on vertical impacts, which allowed us to model the problem in two dimensions. Three-dimensional lagrangian methods such as SPH can investigate non-vertical impacts, but at the expense of lower spatial resolution. Our high-resolution study of vertical impacts is thus complementary to the investigation of oblique impacts reported elsewhere in this issue[16]. A comparison of vertical impacts modelled with both Zeus and SPH suggests that the latter tends to underestimate the volume of crust excavated, primarily because of its limited resolution (see Supplementary Information). In future, a combination of both methods may be required to further address the origin of the dichotomy.

Geological tests of the giant-impact hypothesis have proved inconclusive[17], partly because of later erosion[18] and sedimentation[19], or modification by post-impact slumping[20]. The inferred crustal thickness distribution is less likely to have been modified, except by later impacts or volcanism around Tharsis[21]; lateral crustal flow appears to have been limited[22]. We therefore compared our crustal excavation results with the observed crustal thickness distribution (Fig. 1; see Methods).

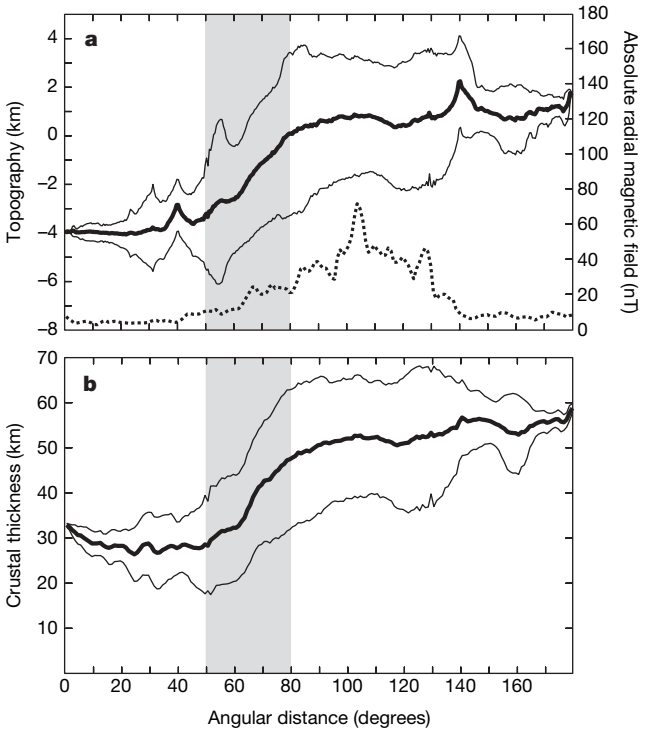


Figure 1 | Averaged crustal properties as a function of angular distance from impact centre (from ref. 6) at 170° E, 50° N. a, Topography (solid lines) from ref. 1 and absolute radial magnetic field (dotted line) at 200-km altitude from ref. 11. **b**, Crustal thickness from ref. 2. Bold lines denote mean values; thin lines denote ±1 s.d.; shaded region denotes basin radius, estimated on the basis of the crustal thickness data to be 50–80° (3,000–4,750 km; compare with ref. 6).

[1]Department of Earth and Planetary Sciences, University of California, Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA. [2]Astronomy Unit, Queen Mary, University of London, London E1 4NS, UK.

Impact melting occurs because of decompression following the initial shock, and depends on the peak shock pressure reached[23,24]. Numerical simulations[25] show that the total mass of melt generated is well approximated by $5U/C^2$, where U is the kinetic energy of the impactor and C is the sound speed (7.8 km s^{-1} for anorthosite at high pressure[23]). The volume of crust missing from the northern lowlands is about 1.5×10^9 km^3; the kinetic energy estimated in ref. 6 implies a melt volume of roughly double that value. To investigate this result, we therefore also calculated the volume of melt generated in our simulations using a second set of model runs in which tracer particles were distributed throughout the crust and mantle (see Methods). We excluded material in an expanded state (density $<$1,000 kg m^{-3}) to avoid counting melted material in the ejecta plume, which does not generally re-impact within the excavated cavity (see Fig. 2).

Figure 2 shows the evolution of a typical impact model. Figure 2a shows a snapshot near the time of maximum transient crater depth. The shock wave has already penetrated into the core, and a region of unexpanded, shock-melted material (grey shading) has developed. However, the bulk of the crust remains undisturbed. At later times (Fig. 2b–d) the ejecta curtain propagates laterally, and in Fig. 2c focusing of a shock wave at the antipode results in crustal disruption and a second, antipodal region of melt generation (see below). Although the ejecta curtain appears to be 'folding back' on itself, this is an artefact which arises because only crustal particles are being plotted. The ejecta actually moves outwards from the impact site in a coherent fashion, with mantle particles forming the core of the apparent fold (see Supplementary Fig. 6).

The width of the excavated region changes little between Fig. 2c and Fig. 2d. The evolution of both the excavation diameter and the percentage of crustal particles escaping Mars are shown in Fig. 2e and demonstrate that constant final values (6,000 km and 2.8%, respectively) are reached at a time of about 4,000 s. Figure 2f shows the final distribution of crustal particles and demonstrates the angular extent of the region from which the crust has been completely stripped. There is an increase in crustal thickness adjacent to the excavated region, a roughly uniform zone in the middle and a ~20°-wide excavated zone at the antipode. The uniform zone, but not the thickened region, is in agreement with the observations (Fig. 1). Because of the two-dimensional nature of the model used, the degree of excavation at the antipode is overestimated (see below) and we can only generate axially symmetric basins (compare with ref. 16).

The total melt volume produced during the impact is 6×10^8 km^3. Ascent and lateral drainage of this melt, although geologically rapid[26], is likely to be slow in comparison with the immediate post-impact

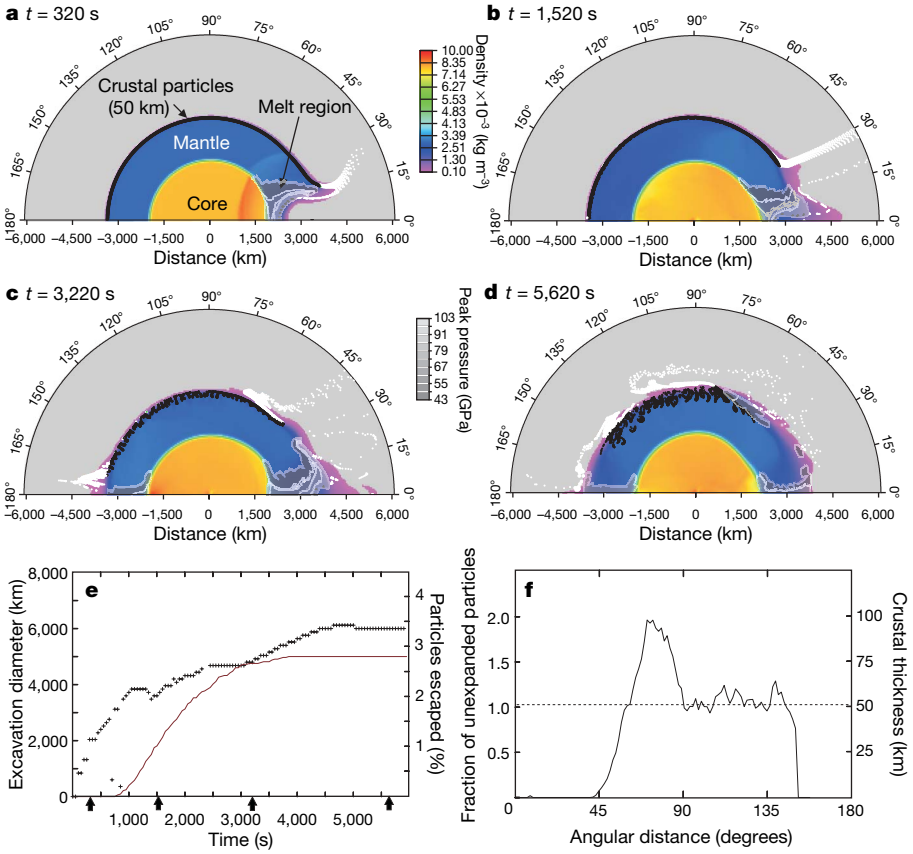


Figure 2 | Results from typical axially symmetric impact simulation using Zeus hydrocode[8,9]. We used 300 evenly spaced grid points in the azimuthal direction and 200 grid points in the radial (r) direction, with a minimum spacing of 25 km for 2,000 $< r <$ 5,000 km and a progressively coarser radial spacing elsewhere (see Supplementary Information). The anorthosite impactor has a radius of 320 km, a vertical speed of 14 km s^{-1} and a kinetic energy of 3.95×10^{28} J. Mars has a radius of 3,400 km and consists of an anorthosite mantle (density 2,940 kg m^{-3}) overlying an iron core of radius 2,000 km and density 7,800 kg m^{-3}; material properties and Tillotson equation of state were taken from ref. 23. The gravity field varies radially but not in time. The model was allowed to equilibrate before impact. **a**–**d**, Snapshots of the resulting density field as a function of time t after impact. Densities $<$100 kg m^{-3} are not plotted. Passive tracer particles representing 50-km-thick crust are white for density $<$1,000 kg m^{-3} and black for density $>$1,000 kg m^{-3}. Grey overlay regions with white contours are areas where unexpanded tracer particles experience peak pressures in excess of the melting threshold (43 GPa; see text). The apparent 'folding back' of the ejecta in **c** and **d** is an artefact caused by plotting only crustal particles (see Supplementary Information). **e**, Time evolution of excavation diameter (crosses) and fraction of crustal particles escaping from Mars (solid line). The excavation diameter was calculated by tracking the fraction of particles in each angular bin (width 60 km) with densities $>$1,000 kg m^{-3}, relative to the original number of particles (see text). Anomalously low points arose owing to failure of the edge-detection algorithm. The arrows denote snapshots shown in **a**–**d**. **f**, Fraction of crustal particles having densities $>$1,000 kg m^{-3} in each angular bin as a function of angular distance at the end of the simulation, calculated using an 11-point moving average. Dotted line denotes initial particle distribution.

slumping and crater rebound[23]. However, ascent will be fast in comparison with the cooling time of the mantle, resulting in an isostatically compensated feature, as observed[2]. The melt will fill in the topographic depression due to the excavated crust. Melt extraction is unlikely to reach 100%, and our adoption of an anorthosite mantle results in conservatively high melt volume estimates (see Supplementary Information). If melt extraction were 100% efficient, a lowland crustal thickness of 21 km (somewhat smaller than that observed) and a topographic low of about 4 km would be generated. A thicker crust and a less pronounced topographic low would result from either a slightly more energetic impactor than that shown in Fig. 2, or subsequent lowland volcanism. Our two-dimensional approach is limited by its inability to model non-vertical impacts. However, because both melt volume and transient crater diameter scale with impact angle in a similar fashion[24], we anticipate that moderately oblique impacts will generate results similar to those discussed here (compare with ref. 16).

Figure 3 summarizes the final excavation radii and mantle melt masses produced for our 28 model runs. Excavating a basin 3,000–4,750 km in radius requires an impact energy of 3×10^{28}–3×10^{29} J, which is comparable to the original estimate of ref. 6. The excavated radii (Fig. 3a) lie close to the line determined by the simple energy-scaling argument used in ref. 6. Except for the most energetic impactors, the corresponding melt masses scale very closely with kinetic energy (Fig. 3b), in agreement with ref. 25. Several model runs with impact energies in the range 3×10^{28}–1×10^{29} J generate a cavity of the correct radius without generating too much melt.

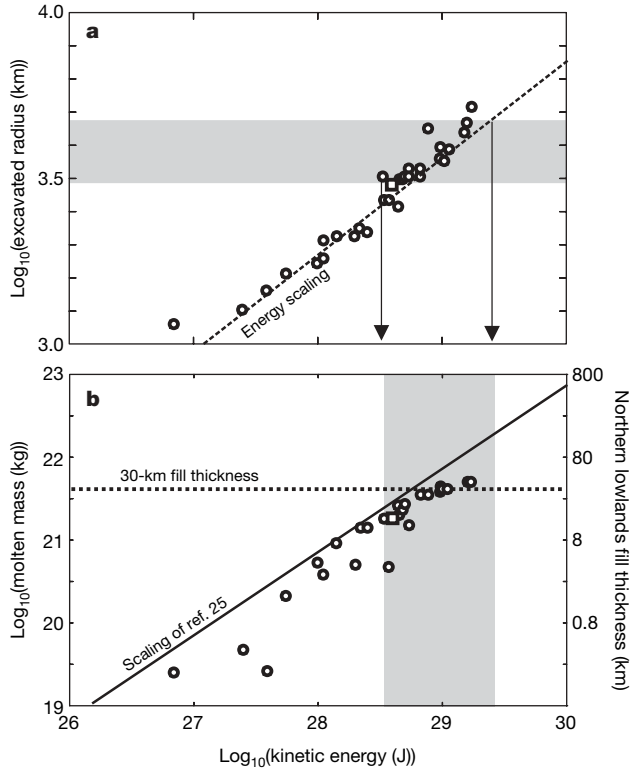


Figure 3 | Impact crustal excavation radius and melt production. **a**, Excavation radius calculated using the method described in Fig. 2 legend, as a function of impact kinetic energy. The shaded region denotes the observed radius of the dichotomy (3,000–4,750 km; Fig. 1) and the arrows give the corresponding range of kinetic energies. The square denotes the model shown in Fig. 2. The dashed line represents the theoretical transient crater radius using the energy-scaling method given in ref. 6. **b**, Melt mass in unexpanded material, calculated using the method described in the text, as a function of kinetic energy. The shaded region gives the approximate energy range permitted by the results in **a**. The dotted line represents the fill thickness required to generate the observed crustal thickness of the northern lowlands (Fig. 1), assuming 100% melt extraction efficiency, a cylindrical basin radius of 3,850 km and a melt density of 2,700 kg m^{-3}. The solid line shows the total melt mass calculated using the scaling of ref. 25 with a sound speed of 7.8 km s^{-1}.

The degree of antipodal excavation seen in our models (Fig. 2) is an overestimate resulting from the perfect focusing caused by the axially symmetric geometry[27]. Three-dimensional SPH simulations show near-surface velocities that are a factor of five to ten lower (see Supplementary Information); the discrepancy is due to a combination of the differing geometries and resolutions used in the two approaches. The SPH antipodal shock pressures (~0.1 GPa) provide a conservative lower bound; they are an order of magnitude too small to cause shock demagnetization[28], but are certainly sufficient to cause fracturing and motion of crustal blocks. Reorientation of previously magnetized crustal material will disrupt any pre-existing long-wavelength, coherently magnetized structures[29] and cause an apparent reduction in magnetization (see, for example, ref. 10) at spacecraft altitudes. This prediction can be tested with low-altitude or surface magnetic field measurements.

A second prediction is that the origin of the northern lowlands crust is fundamentally different from that of the highlands crust. The latter probably formed directly by magma ocean solidification a few tens of millions of years after Mars accreted[12]. The northern lowlands crust, by contrast, probably arose primarily from shock melting in the deep and previously depleted martian mantle (Fig. 2). The northern lowlands crust is thus likely to have a different composition and density to that of the highlands. Although detailed petrological modelling of such a melt source has yet to be carried out, generation of shergottites by deep melting is a possibility[30]. Furthermore, an impact model for the northern lowlands crust is consistent with evidence of a partial melt derived from a depleted source which formed ~100 Myr after accretion[13]. If correct, this observation agrees with crater-counting estimates of the lowlands' age[15], and suggests that the dichotomy-forming impact took place within the same time window as the Moon-forming impact on Earth[14].

METHODS SUMMARY

We determined the radius of the region from which crustal material had been excavated by tracking the fraction f of passive crustal tracer particles remaining within a series of angular bins, relative to the original number. Particles that had densities less than a critical value ρ_c were assumed to be in an expanded state[23] and were not counted; varying ρ_c from 100 to 2,000 kg m^{-3} had negligible effect on the results. The edge of the excavated region was found by establishing where a moving average of f first exceeded a critical value (see Methods). To investigate melting, we carried out a second set of model runs in which tracer particles were distributed throughout the entire crust and mantle. Following ref. 24, we determined the degree of melting for each particle by comparing the maximum pressure experienced with the pressure range over which melting occurs (43–52 GPa for anorthosite at high pressure[23]), and integrated over all the particles to estimate the total melt volume.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 5 December 2007; accepted 14 April 2008.

1. Smith, D. E. *et al.* The global topography of Mars and implications for surface evolution. *Science* **284**, 1495–1503 (1999).
2. Neumann, G. A. *et al.* Crustal structure of Mars from gravity and topography. *J. Geophys. Res.* **109**, E08002 (2004).
3. Watters, T. R., McGovern, P. J. & Irwin, R. P. Hemispheres apart: The crustal dichotomy on Mars. *Annu. Rev. Earth Planet. Sci.* **35**, 621–652 (2007).
4. Wise, D. U., Golombek, M. P. & McGill, G. E. Tectonic evolution of Mars. *J. Geophys. Res.* **84**, 7934–7939 (1979).
5. Zhong, S. J. & Zuber, M. T. Degree-1 mantle convection and the crustal dichotomy on Mars. *Earth Planet. Sci. Lett.* **189**, 75–84 (2001).
6. Wilhelms, D. E. & Squyres, S. W. The Martian hemispheric dichotomy may be due to a giant impact. *Nature* **309**, 138–140 (1984).
7. Frey, H. & Schultz, R. A. Large impact basins and the mega-impact origin for the crustal dichotomy on Mars. *Geophys. Res. Lett.* **15**, 229–232 (1988).
8. Stone, J. M. & Norman, M. L. Zeus-2D – a radiation magnetohydrodynamics code for astrophysical flows in 2 space dimensions. 1. The hydrodynamic algorithms and tests. *Astrophys. J. Suppl. Ser.* **80**, 753–790 (1992).

9. Korycansky, D. G., Zahnle, K. J., &. Mac Low, M. M. High-resolution calculations of asteroid impacts into the Venusian atmosphere. *Icarus* **146,** 387–403 (2000).
10. Beals, C. S., Innes, M. J. S. & Rottenburg, J. A. in *The Moon, Meteorites and Comets* (ed. Middlehurst, B. M.) 235–284 (Univ. Chicago Press, Chicago, 1963).
11. Purucker, M. *et al.* An altitude-normalized magnetic map of Mars and its interpretation. *Geophys. Res. Lett.* **27,** 2449–2452 (2000).
12. Nyquist, L. E. *et al.* Ages and geologic histories of Martian meteorites. *Space Sci. Rev.* **96,** 105–164 (2001).
13. Debaille, V., Brandon, A. D., Yin, Q. Z. & Jacobsen, B. Coupled ^{142}Nd–^{143}Nd evidence for a protracted magma ocean in Mars. *Nature* **450,** 525–528 (2007).
14. Kleine, T., Palme, H., Mezger, M. & Halliday, A. N. Hf-W chronometry of lunar metals and the age and early differentiation of the Moon. *Science* **310,** 1671–1674 (2005).
15. Frey, H. V. Impact constraints on the age and origin of the lowlands of Mars. *Geophys. Res. Lett.* **33,** L08502 (2006).
16. Marinova, M. M., Aharonson, O. & Asphaug, E. Mega-impact formation of the Mars hemispheric dichotomy. *Nature* doi:10.1038/nature07070 (this issue).
17. McGill, G. E. & Squyres, S. W. Origin of the Martian crustal dichotomy – evaluating hypotheses. *Icarus* **93,** 386–393 (1991).
18. Hynek, B. M. & Phillips, R. J. Evidence for extensive denudation of the Martian highlands. *Geology* **29,** 407–410 (2001).
19. Tanaka, K. L. Sedimentary history and mass flow structures of Chryse and Acidalia Planitiae, Mars. *J. Geophys. Res.* **102,** 4131–4149 (1997).
20. Spudis, P. D. *The Geology of Multi-Ring Impact Basins* (Cambridge Univ. Press, Cambridge, UK, 2005).
21. Phillips, R. J. *et al.* Ancient geodynamics and global-scale hydrology on Mars. *Science* **291,** 2587–2591 (2001).
22. Zuber, M. T. *et al.* Internal structure and early thermal evolution of Mars from Mars Global Surveyor topography and gravity. *Science* **287,** 1788–1793 (2000).
23. Melosh, H. J. *Impact Cratering: A Geologic Process* (Oxford Univ. Press, Oxford, UK, 1989).
24. Pierazzo, E. & Melosh, H. J. Melt production in oblique impacts. *Icarus* **145,** 252–261 (2000).
25. O'Keefe, J. D. & Ahrens, T. J. Impact-induced energy partitioning, melting and vaporization on terrestrial planets. *Proc. Lunar Sci. Conf. 8th* **3,** 3357–3374 (1977).
26. Turner, S., Evans, P. & Hawkesworth, C. Ultrafast source-to-surface movement of melt at island arcs from Ra-226-Th-230 systematics. *Science* **292,** 1363–1366 (2001).
27. Bruesch, L. S. & Asphaug, E. Modeling global impact effects on middle-sized icy bodies: applications to Saturn's moons. *Icarus* **168,** 457–466 (2004).
28. Hood, L. L., Richmond, N. C., Pierazzo, E. & Rochette, P. Distribution of crustal magnetic fields on Mars: Shock effects of basin-forming impacts. *Geophys. Res. Lett.* **30,** doi:10.1029/2002GL016657 (2003).
29. Acuna, M. H. *et al.* Global distribution of crustal magnetization discovered by the Mars Global Surveyor MAG/ER experiment. *Science* **284,** 790–793 (1999).
30. Agee, C. B. & Draper, D. S. Experimental constraints on the origin of Martian meteorites and the composition of the Martian mantle. *Earth Planet. Sci. Lett.* **224,** 415–429 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was funded by NASA's MFR programme.

Author Contributions F.N. and S.D.H. carried out the two-dimensional runs and analysed the results, D.G.K. modified the two-dimensional code for the present application and C.B.A. carried out the SPH runs. F.N. conceived the project and wrote the first draft of the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to F.N. (fnimmo@es.ucsc.edu).

METHODS

The model grid consisted of 300 points in the azimuthal direction and 200 in the radial direction. The radial grid consisted of three regions: an inner region ($r < 2{,}000$ km, 30 grid points) in which the distance between successive grid points decreased outwards by a factor of 0.952; a central region (2,000 km $< r <$ 5,000 km, 120 grid points) in which the distance was constant at 25 km; and an outer region (5,000 km $< r <$ 9,000 km, 50 grid points) in which the distance between successive grid points increased outwards by a factor of 1.04.

The model consisted of an iron core (radius 2,000 km), an anorthosite mantle (outer radius 3,400 km) and a thin external atmosphere (surface density 10^{-2} kg m^{-3}, scale height 11 km). Solid material used the Tillotson equation of state[23]; the atmosphere was assumed to be an ideal gas. The interior gravitational acceleration was calculated assuming the initial two-layer, constant-density structure, whereas the exterior gravity falls off as $1/r^2$. We did not recalculate gravitational accelerations at subsequent time steps, but instead assumed them to remain at the initial values.

We initialized the target in an isothermal state with each layer having constant density. We then allowed it to equilibrate for 3,380 s before an impact was imposed, to allow the target material to reach hydrostatic equilibrium. Impact parameters for individual runs are tabulated in Supplementary Table 1, and azimuthally averaged initial values of energy density, pressure and density for selected radial distances from the centre of the planet are tabulated in Supplementary Table 2.

We calculated the evolution of the excavated crustal diameter by determining the fraction f of unexpanded (density $>1{,}000$ kg m^{-3}) particles left within each angular bin of width 60 km. The edge of the excavated region was set to bin index j when the following criteria were satisfied: $f_{j-2} = 0$, $f_{j-1} < 0.5$, $f_{j+2} > 0.5$, $f_{j+3} > 0.5$. This particular set of criteria was found to be robust when compared with visual inspection of the excavated cavity diameter.

We carried out two sets of runs that differed only in the distribution of passive tracer particles. The first set used tracer particles separated by 5 km in the vertical direction and 20 km (at the surface) in the horizontal direction, in the top 50 km of the solid region only. The second set used tracer particles separated by 10 km and 60 km (at the surface) in the vertical and horizontal directions, respectively, over the entire crust and mantle of the model. We used the first set of runs to investigate crustal excavation, and the second set to investigate melt production.

To calculate the total melt volume, we first calculated the maximum shock pressure experienced by each particle, P_{max}, and then calculated the resulting melt fraction from $(P_{max} - P_{im})/(P_{cm} - P_{im})$, where $P_{im} = 43$ GPa and $P_{cm} = 52$ GPa are respectively the pressures for incipient and complete melting[23], and the melt fraction saturates at a value of one. Only unexpanded particles (density $>1{,}000$ kg m^{-3}) were counted.

LETTERS

A BCS-like gap in the superconductor $SmFeAsO_{0.85}F_{0.15}$

T. Y. Chen[1], Z. Tesanovic[1], R. H. Liu[2], X. H. Chen[2] & C. L. Chien[1]

Since the discovery of superconductivity in the high-transition-temperature (high-T_c) copper oxides two decades ago, it has been firmly established that the CuO_2 plane is essential for superconductivity and gives rise to a host of other very unusual properties. A new family of superconductors with the general composition of $LaFeAsO_{1-x}F_x$ has recently been discovered[1–8] and the conspicuous lack of the CuO_2 planes raises the tantalizing question of a different pairing mechanism in these oxypnictides. The superconducting gap (its magnitude, structure, and temperature dependence) is intimately related to pairing. Here we report the observation of a single gap in the superconductor $SmFeAsO_{0.85}F_{0.15}$ with $T_c = 42$ K as measured by Andreev spectroscopy. The gap value of $2\Delta = 13.34 \pm 0.3$ meV gives $2\Delta/k_BT_c = 3.68$ (where k_B is the Boltzmann constant), close to the Bardeen–Cooper–Schrieffer (BCS) prediction of 3.53. The gap decreases with temperature and vanishes at T_c in a manner consistent with the BCS prediction, but dramatically different from that of the pseudogap behaviour in the copper oxide superconductors. Our results clearly indicate a nodeless gap order parameter, which is nearly isotropic in size across different sections of the Fermi surface, and are not compatible with models involving antiferromagnetic fluctuations, strong correlations, the *t-J* model, and the like, originally designed for the high-T_c copper oxides.

The $LaFeAsO_{1-x}F_x$ superconductors have a tetragonal structure with lattice parameters of $a \approx 4.03$ Å and $c \approx 8.73$ Å, consisting of alternating layers of quasi-two-dimensional puckered LaO and FeAs planes along the *c* axis. In the non-superconducting parent compound of LaFeAsO, the Fe magnetic moments exhibit spin-density-wave antiferromagnetic ordering below about 150 K (ref. 9). By replacing O with F, the FeAs layers can be doped with electrons, resulting in the decrease of the Néel temperature, T_N, and the emergence of superconductivity at a doping level of $x \approx 0.15$–0.2 in $LaFeAsO_{1-x}F_x$ (ref. 7).

An essential physical quantity in any superconductor is the superconducting gap 2Δ, whose value and structure are intimately related to the pairing mechanism. Band structure calculations have revealed disconnected Fermi surfaces in $LaFeAsO_{1-x}F_x$ (ref. 10), thus raising the prospect of two superconducting gaps, a situation previously encountered in MgB_2 (ref. 11). In this work, we report measurements of the superconducting gap of $SmFeAsO_{0.85}F_{0.15}$ by Andreev spectroscopy, one of the relatively few methods by which the superconducting gap can be determined. The fabrication and superconducting properties (with $T_c \approx 42$ K) of the polycrystalline $SmFeAsO_{0.85}F_{0.15}$ sample have been described elsewhere[3]. We have independently confirmed the value of T_c by resistance and Meissner effect measurements. For the Andreev spectroscopy measurements, we used Au tips in contact with a sample of $SmFeAsO_{0.85}F_{0.15}$ about 5 mm in size. The measurements of differential conductance $dI/dV(V) = G(V)$, where I is current, V is bias voltage across the contact, and G is conductance, were carried out by varying V in a variable-temperature cryostat; measurements were made mostly without applied magnetic field, but some were performed with a magnetic field of up to 9 T. Negative V is defined as being when electrons are injected from the tip into the superconductor.

At a normal metal/superconductor interface, the injected current at a bias voltage within the gap must first be converted into a supercurrent consisting of Cooper pairs of electrons with opposite spins. This can be accomplished by having the injected electron from one spin band accompanied by another electron from the opposite spin band. This is the well-known Andreev reflection process, which is equivalent to reflecting a hole back into the metal, thus doubling the conductance within the superconducting gap. Andreev spectroscopy provides a sensitive and quantitative measurement of the gap structure of superconductors[11,12]. The measured conductance $G(V)$ can be quantitatively analysed using well-developed theoretical models, such as the modified Blonder–Tinkham–Klapwijk (BTK) model[12–14], which include effects due to thermal broadening and the less than ideal contacts often encountered.

We first illustrate the Andreev spectroscopy measurements at 4.52 K of two well-known superconductors. The normalized conductance $G(V)/G_n$, where G_n is the normal state conductance, of Nb (Fig. 1a) exhibits two peaks, indicating a single gap with a value of $2\Delta = 2.84$ meV, whereas that of MgB_2 displays two gaps with values of $\Delta_S = 2.75$ meV and $\Delta_L = 6.75$ meV (Fig. 1b). These results show that the Andreev spectroscopy can accurately determine the superconducting gap, or even multiple gaps only a few meV apart, using polycrystalline samples of Nb and MgB_2.

One representative conductance result of an Au/$SmFeAsO_{0.85}F_{0.15}$ contact at 4.52 K within the bias voltage range of ± 30 mV is shown in Fig. 1c. There are two peaks with a separation of about 13.2 mV, indicating a single gap, whose apparent value is close to $2\Delta = 13.6$ meV obtained from the BTK analysis, which gives an excellent description of the conductance results within ± 30 mV, as shown by the red curve through the data points. However, to ensure confidence in this gap value for $SmFeAsO_{0.85}F_{0.15}$, it is essential to examine the conductance results of all the contacts and address the unexpected features. We have made conductance measurements at 4.52 K of about 15 contacts extending to ± 100 mV, with and without a magnetic field. Some of these results are shown in Fig. 2, arranged top-to-bottom in decreasing value of the contact resistance R, which scales inversely with the contact size. The actual values of the contact size can in principle be obtained using the Wexler formula[15] with the knowledge of the electrical resistivity, which for polycrystalline $SmFeAsO_{0.85}F_{0.15}$ pellet samples unfortunately remains quite tenuous. All the measured conductance $G(V)$ results show a pronounced symmetric structure within $\pm$ 25 mV, while some contain additional small but complex structures. Despite all the extra structures, however, it is important to note that all the conductance curves contain

[1]Department of Physics and Astronomy, Johns Hopkins University, Baltimore, Maryland 21218, USA. [2]Hefei National Laboratory for Physical Sciences at Microscale and Department of Physics, University of Science and Technology of China, Hefei, Anhui 230026, China.

the same feature—two peaks with an apparent separation of about 13.2 mV, as indicated by the vertical dashed lines (Fig. 2). It is therefore conclusively established that 2Δ of the $SmFeAsO_{0.85}F_{0.15}$ sample is set by this value. The average of the best-fit values from all the analysed Andreev spectra at 4.52 K is $2\Delta = 13.34 \pm 0.3$ meV. Together with $T_c = 42$ K, we obtain $2\Delta/k_BT_c = 3.68$, which is close to, although slightly higher than, the well-known universal value of $2\Delta/k_BT_c = 3.53$ within the BCS theory. Our method of identifying the superconducting gap is further corroborated by its temperature dependence, as described below.

It is often observed in the $G(V)/G_n$ of normal metal/superconductor contacts that a spike appears at $V = 0$, known as the zero bias anomaly (ZBA). The ZBA can have such a high intensity (>2) that it overwhelms the entire conductance spectrum. We have observed the ZBA, and our data show its systematic emergence (Fig. 2). As the contact size becomes larger (hence a smaller contact resistance R), the ZBA increases in intensity. At $R = 42\,\Omega$ (Fig. 2e), the ZBA has protruded through the gap structure. At $R = 4.3\,\Omega$ (Fig. 2f), the ZBA is so large (exceeding 7; note the different y-axis scale) that it dwarfs the gap structure. The ZBA has been observed in copper oxide superconductors with d-wave pairing[16] as well as in low-T_c s-wave superconductors such as Al (ref. 17) and Nb (ref. 18). Thus the ZBA is not exclusive to d-wave pairing. It is clear, however, that the ZBA is intimately related to the superconductivity, since a ZBA does not appear at a temperature above T_c, nor in contacts of normal metals. The ZBA is also related to the contact size, hence the contact resistance, as illustrated in Fig. 2. Consequently, conductance results containing a dominant ZBA contribution (for example, Fig. 2f) would render the extraction of the gap structure fruitless. We have found that a 9 T magnetic field at 4.52 K has very little effect on both the extraneous structure and the superconducting gap feature, as shown in Fig. 2d, f, in sharp contrast to the claim that the ZBA can be substantially reduced by a magnetic field of only a few teslas (ref. 19).

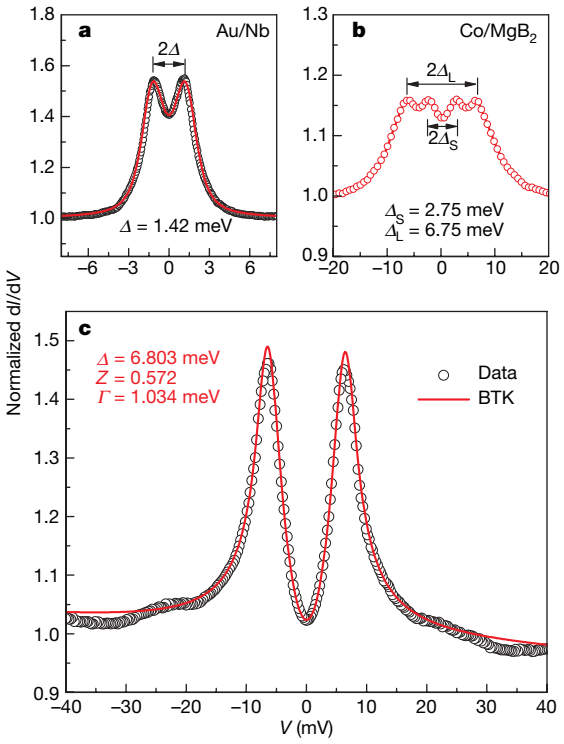


Figure 1 | Representative Andreev spectra at 4.52 K. **a**, An Au/Nb contact shows one gap. **b**, A Co/MgB_2 contact shows two distinct gaps. **c**, An Au/$SmFeAsO_{0.85}F_{0.15}$ contact shows one gap. Open circles are the experimental data, and the solid curves in **a** and **c** are the best-fit results to the modified BTK model[14] with parameters shown. Details of experimental procedure and data analysis are given in Supplementary Information.

In addition to the main gap, all the conductance curves also show extra features, some extending to ± 100 mV. However, although the actual causes are not known, these features appear to be related to superconductivity as they disappear at higher temperatures. It is worth mentioning that the conductance curves of high-quality single-crystal copper oxide superconductors, such as $La_{2-x}Sr_xCuO_4$, also show rather complex additional features outside the gap structure[20]. Another unusual feature shown in Fig. 2 is the asymmetric background, always with $G(-V) > G(V)$. However, the asymmetrical background remains at temperatures much higher than T_c, so its origin appears to be unrelated to superconductivity; it may perhaps be due to the large mismatch of conductivity between metal and oxide conductors. We note that most results of scanning tunnelling microscopy investigations of copper oxide superconductors display an acute asymmetrical background[21]. In contrast, contacts with low-T_c metallic superconductors, such as Nb and MgB_2, do not exhibit these features. The conductance curves shown in Fig. 1a, b are flat, extending to very large voltages, and show the expected symmetry of $G(V) = G(-V)$.

The temperature dependence of the superconducting gap, in addition to the gap value, is also of great importance. We have determined the temperature dependence of the gap of $SmFeAsO_{0.85}F_{0.15}$ by varying the temperature of one contact, and measured more than 70 conductance curves at various temperatures from 4.5 K to 60 K. These results are summarized in Fig. 3a. At low temperatures, the gap structure is resolved and has a high intensity. The separation of

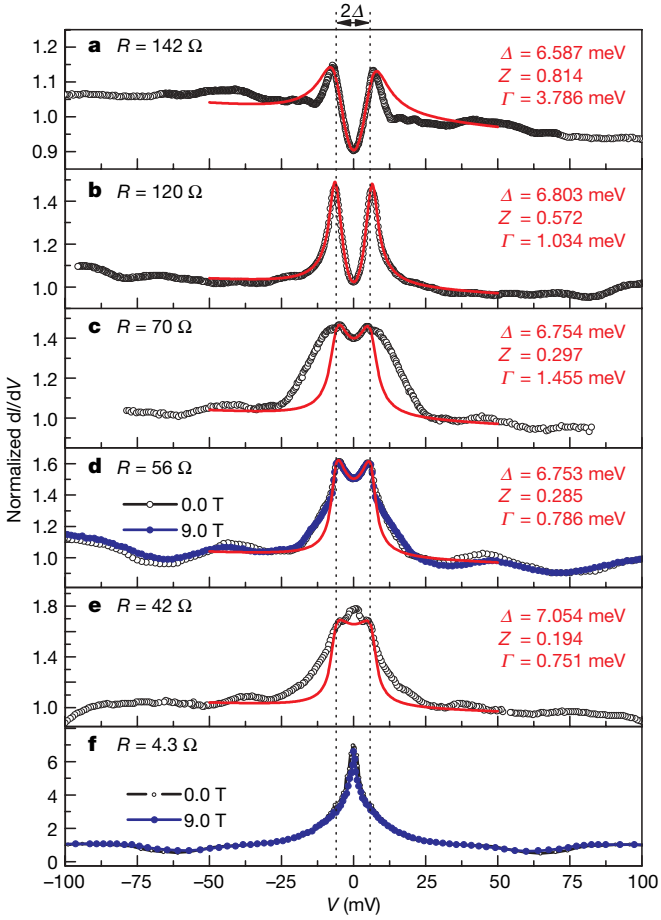


Figure 2 | Andreev spectra of Au/$SmFeAsO_{0.85}F_{0.15}$ point contacts at 4.52 K with various contact resistances. **a–f**, Spectra are arranged in order of decreasing contact resistance R; open circles are the experimental data, and red solid curves are the best fit to the modified BTK model with the parameters and contact resistance listed in each panel. The vertical dashed lines are at ± 6.6 meV to indicate the common features. The blue curve in **d** and **f** shows results taken in an external magnetic field of $H = 9$ T.

the two peaks, and hence the value of the gap, decreases with increasing temperature, and evolves into a single unresolved peak with decreasing width. The intensity of the central peak also decreases with increasing temperature and vanishes at T_c; it is not detectable at temperatures above T_c. The gap value, obtained from analyses of the conductance curves, shows a temperature dependence (Fig. 3b). This measured temperature dependence of Δ is close to the BCS prediction, which is shown by the dashed curve.

Although we have used a polycrystalline $SmFeAsO_{0.85}F_{0.15}$ sample, the measured result of a single gap may not be the result of averaging. Owing to the small contact size (especially those with large contact resistances), the gap value is most often measured from a single grain. As numerous contacts have yielded essentially the same gap value at a given temperature, the gap structure in $SmFeAsO_{0.85}F_{0.15}$ is likely to be nearly isotropic. Ultimately, key questions such as the possibility of an anisotropic gap and the existence of a nodal gap associated with *d*- or *p*-wave pairing can only be addressed using single-crystal specimens. Our results also show that if multiple gaps exist in $SmFeAsO_{0.85}F_{0.15}$, their values must be very close to each other, as only a single sharply defined gap, without any discernable structure, has been observed (Fig. 1c).

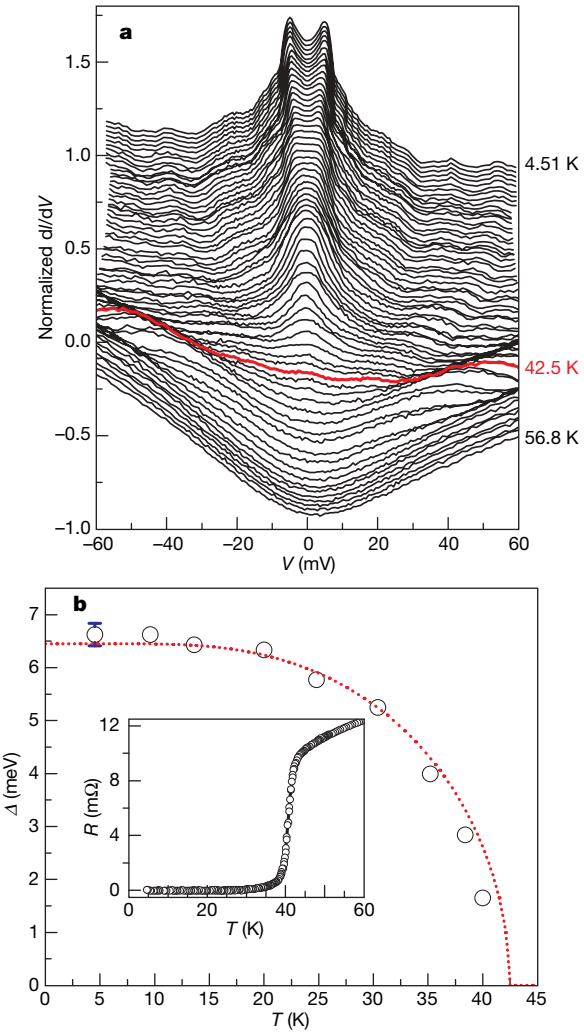


Figure 3 | Temperature dependence of the gap. a, Andreev spectra from 4.51 K to 56.8 K of an Au/$SmFeAsO_{0.85}F_{0.15}$ contact taken every ~0.8 K, showing the decrease of the gap value and the eventual disappearance of the gap structure at $T_c \approx 42$ K. **b**, Main panel, temperature dependence of the gap value (open circles) obtained from the modified BTK fit of $SmFeAsO_{0.85}F_{0.15}$ is close to the BCS theory prediction (dotted curve). Error bars at 4.51 K show ±1 s.e.m. from the average of 15 spectra for different contacts. Inset, the resistive transition at T_c.

The gap value and its temperature dependence in $SmFeAsO_{0.85}F_{0.15}$ are dramatically different from those found in the pseudogap regime of high-T_c copper oxide superconductors. The pseudogap of the copper oxide superconductors is large, much larger than the value from the BCS prediction, and most importantly, it is essentially temperature independent[21]. Furthermore, not only does the tunnelling pseudogap persist at the same value, it can even be resolved at temperatures much higher than T_c, sometimes as high as $3T_c$. Even in some Andreev reflection experiments, where the pseudogap gap disappears closer to T_c, its $T = 0$ value is far in excess of BCS values for $2\Delta/k_BT_c = 3.53$ (*s*-wave) and $2\Delta/k_BT_c = 4.28$ (*d*-wave). These unique characteristics of the pseudogap in the copper oxide superconductors are absent in $SmFeAsO_{0.85}F_{0.15}$.

The gap value Δ of the $SmFeAsO_{0.85}F_{0.15}$ superconductor and its temperature dependence are close to the BCS predictions. However, the exceptionally high T_c of 40–55 K seems to have exceeded the limit of the electron–phonon coupling. The actual mechanism of superconductivity in $SmFeAsO_{0.85}F_{0.15}$ has yet to be determined, and there are already a number of theoretical proposals. Our results provide clear evidence of two central aspects of the pairing mechanism in superconducting $SmFeAsO_{0.85}F_{0.15}$: it is a single-gap superconductor, its gap exhibiting a BCS-type temperature dependence, and there are no signatures of the pseudogap structure that features prominently in the copper oxide superconductors. Furthermore, we do not see evidence of nodal structure in the gap, and the observed $2\Delta/k_BT_c$ is notably less than the weak-coupling *d*- and *p*-wave values. Our results therefore imply a nodeless, BCS-type gap, and set strong constraints on the ongoing theoretical research.

Received 23 April; accepted 5 May 2008.
Published online 4 June 2008.

1. Kamihara, Y., Watanabe, T., Hirano, M. & Hosono, H. Iron-based layered superconductor $La[O_{1-x}F_x]FeAs_x$ (x = 0.05–0.12) with *Tc* = 26 K. *J. Am. Chem. Soc.* **130,** 3296–3297 (2008).
2. Wen, H. H., Mu, G., Fang, L., Yang, H. & Zhu, X. Y. Superconductivity at 25 K in hole doped $La_{1-x}Sr_xOFeAs$. Preprint at ⟨http://arXiv.org/abs/0803.3021⟩ (2008).
3. Chen, X. H. *et al.* Superconductivity at 43 K in samarium-arsenide oxides $SmFeAsO_{1-x}F_x$. Preprint at ⟨http://arXiv.org/abs/0803.3603⟩ (2008).
4. Chen, G. F. *et al.* Superconductivity at 41 K and its competition with spin-density-wave instability in layered $CeO_{1-x}F_xFeAs$. Preprint at ⟨http://arXiv.org/abs/0803.3790⟩ (2008).
5. Ren, Z. A. *et al.* Superconductivity at 52 K in iron-based F-doped layered quaternary compound $Pr[O_{1-x}F_x]FeAs$. Preprint at ⟨http://arXiv.org/abs/0803.4283⟩ (2008).
6. Ren, Z. A. *et al.* Superconductivity at 55 K in iron-based F-doped layered quaternary compound $Sm[O_{1-x}F_x]FeAs$. Preprint at ⟨http://arXiv.org/abs/0804.2053⟩ (2008).
7. Liu, R. H. *et al.* Phase diagram and quantum critical point in newly discovered superconductors: $SmO_{1-x}F_xFeAs$. Preprint at ⟨http://arXiv.org/abs/0804.2105⟩ (2008).
8. Ren, Z. A. *et al.* Novel superconductivity and phase diagram in the iron-based arsenic-oxides $ReFeAsO_{1-delta}$ (Re = rare earth metal) without F-doping. Preprint at ⟨http://arXiv.org/abs/0804.2582⟩ (2008).
9. de la Cruz, C. *et al.* Magnetic order versus superconductivity in the iron-based layered $La(O_{1-x}F_x)FeAs$ systems. Preprint at ⟨http://arXiv.org/abs/0804.0795⟩ (2008).
10. Mazin, I. I., Singh, D. J., Johannes, M. D. & Du, M. H. Unconventional sign-reversing superconductivity in $LaFeAsO_{1-x}F_x$. Preprint at ⟨http://arXiv.org/abs/0803.2740⟩ (2008).
11. Szabó, P. *et al.* Evidence for two superconducting energy gaps in MgB_2 by point-contact spectroscopy. *Phys. Rev. Lett.* **87,** 137005 (2001).
12. Strijkers, G. J., Ji, Y., Yang, F. Y., Chien, C. L. & Byers, J. M. Andreev reflections at metal/superconductor point contacts: Measurement and analysis. *Phys. Rev. B* **63,** 104510 (2001).
13. Blonder, G. E., Tinkham, M. & Klapwijk, T. M. Transition from metallic to tunneling regimes in superconducting microconstrictions: Excess current, charge imbalance, and supercurrent conversion. *Phys. Rev. B* **25,** 4515–4532 (1982).
14. Plecenik, A., Grajcar, M., Beňačka, Š., Seidel, P. & Pfuch, A. Finite-quasiparticle-lifetime effects in the differential conductance of $Bi_2Sr_2CaCu_2O_y$/Au junctions. *Phys. Rev. B* **49,** 10016–10019 (1994).
15. Wexler, G. The size effect and the non-local Boltzmann transport equation in orifice and disk geometry. *Proc. Phys. Soc.* **89,** 927–941 (1966).
16. Deutscher, G. Andreev-Saint-James reflections: A probe of cuprate superconductors. *Rev. Mod. Phys.* **77,** 109–135 (2005).

17. Li, G. *et al.* Tunneling spectroscopy in AlNiCo decagonal quasicrystals. *Phys. Rev. Lett.* **82,** 1229–1232 (1999).
18. Xiong, P., Xiao, G. & Laibowitz, R. B. Subgap and above-gap differential resistance anomalies in superconductor-normal-metal microjunctions. *Phys. Rev. Lett.* **71,** 1907–1910 (1993).
19. Shan, L. *et al.* Unconventional pairing symmetry in iron-based layered superconductor $LaO_{0.9}F_{0.1-delta}FeAs$ revealed by point-contact spectroscopy measurements. Preprint at ⟨http://arXiv.org/abs/0803.2405⟩ (2008).
20. Achsaf, N., Deutscher, G., Revcolevschi, A. & Okuya, M. in *Coherence in High Temperature Superconductors* (eds Deutscher, G. & Revcolevschi, A.) 428–442 (World Scientific, Singapore, 1996).
21. Fischer, O., Kugler, M., Maggio-Aprile, I. & Bertod, C. Scanning tunneling spectroscopy of high-temperature superconductors. *Rev. Mod. Phys.* **79,** 353–419 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the US National Science Foundation and the Natural Science Foundation of China.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.L.C. (clc@pha.jhu.edu).

Vol 453|26 June 2008|doi:10.1038/nature07046

LETTERS

The total synthesis of (−)-cyanthiwigin F by means of double catalytic enantioselective alkylation

John A. Enquist, Jr[1] & Brian M. Stoltz[1]

Double catalytic enantioselective transformations are powerful synthetic methods that can facilitate the construction of stereochemically complex molecules in a single operation[1,2]. In addition to generating two or more stereocentres in a single reaction, multiple asymmetric reactions also impart increased enantiomeric excess to the final product in comparison with the analogous single transformation[3–6]. Furthermore, multiple asymmetric operations have the potential to independently construct several stereocentres at remote points within the same molecular scaffold, rather than relying on pre-existing chiral centres that are proximal to the reactive site[1]. Despite the inherent benefits of multiple catalytic enantioselective reactions, their application to natural product total synthesis remains largely underutilized[2]. Here we report the use of a double stereoablative[7] enantioselective alkylation reaction in a concise synthesis of the marine diterpenoid (−)-cyanthiwigin F (ref. 8). By employing a technique for independent, selective formation of two stereocentres in a single stereoconvergent operation, we demonstrate that a complicated mixture of racemic and meso diastereomers may be smoothly converted to a synthetically useful intermediate with exceptional enantiomeric excess. The stereochemical information generated by means of this catalytic transformation facilitates the easy and rapid completion of the total synthesis of this marine natural product.

Originally isolated from the sea sponge *Myrmekioderma styx*, cyanthiwigin F (**1**; see Fig. 1) is one of 30 known cyanthiwigin natural products, all of which belong to a larger class of bioactive molecules known as the cyathins[8–10]. The cyathins display a wide range of biological properties including antimicrobial activity, antineoplastic action, stimulation of nerve growth factor synthesis, and κ-opioid receptor agonism[9,11]. Cyanthiwigin F itself exhibits cytotoxic activity against human primary tumour cells (with a half-maximal inhibitory concentration of 3.1 μg ml^{-1})[8]. A skeletal representation of the cyathins (that is, **2**; see Fig. 1) depicts the 20-carbon tricyclic core bearing two all-carbon quaternary stereocentres at the A–B and B–C ring junctures. Synthetic control of the relative and absolute stereochemistries of these quaternary centres along with the construction of the tricyclic framework of the cyathins represent significant challenges to their laboratory preparation[12,13]. Despite their structural and biological significance, only two of the 30 cyanthiwigin molecules isolated so far have been accessed by total synthesis[14,15]. Owing to the promising activity and low natural abundance of these molecules, total synthesis remains an important technique in the production of these compounds in quantities sufficient for further study. Moreover, a succinct and general approach to such a synthetic effort may provide a strategic opportunity for the preparation of many other members of the cyanthiwigin family.

Our retrosynthetic approach to cyanthiwigin F focused on initial construction of the central B ring, with the intention of rapidly establishing both of the all-carbon quaternary stereocentres present at the ring junctures of the natural product (Fig. 1). To this end, we envisioned late-stage construction of the five-member A ring, leading back to bicyclic ketone **3**. We anticipated that this bicyclic structure would be accessible by means of ring-closing metathesis of cyclohexanone **4**, which itself could be prepared via Negishi cross-coupling of vinyl triflate **5**. Critically, we envisioned that triflate **5** could be obtained from monoanionic desymmetrization of cyclohexanedione **6** (ref. 16). We recently developed a powerful methodology for the construction of all-carbon quaternary stereocentres using a chiral palladium catalyst to mediate the asymmetric alkylation of *in*

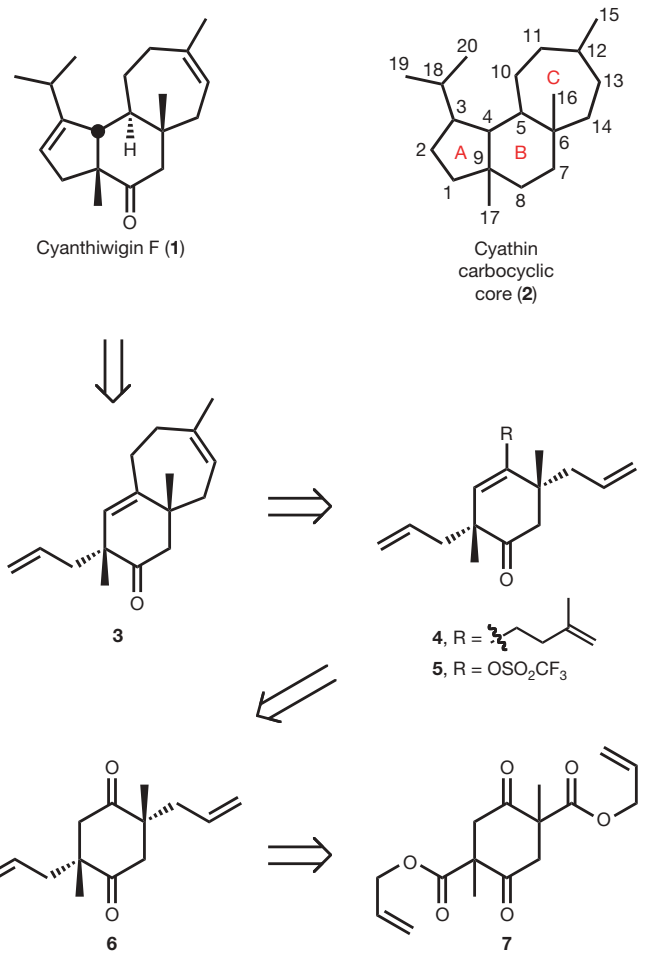


Figure 1 | Structure and retrosynthesis of cyanthiwigin F.

[1]The Arnold and Mabel Beckman Laboratories of Chemical Synthesis, Division of Chemistry and Chemical Engineering, California Institute of Technology, 1200 East California Boulevard, MC164-30, Pasadena, California 91125, USA.

situ-generated cyclic ketone enolates[17,18]. Using this technology, we anticipated that cyclohexanedione **6** could be accessed through enantioselective allylation of bis(β-ketoester) **7**, which could be generated from diallyl succinate (**8**; see Fig. 2a). Indeed, in a model substrate, a double asymmetric alkylation using our protocol proved successful[18].

A key challenge presented by the current system is that it represents the first Pd-catalysed double stereoablative alkylation event involving two different ketones contained within a single ring to generate two all-carbon quaternary stereocentres. Additionally, the current substrate is unique in that a mixture of diastereomers and enantiomers would be converted to a single stereoisomer of product. This transformation would be particularly noteworthy because non-selective catalytic ablation of two quaternary centres in the substrate would precede highly enantio- and diastereoselective construction of two central quaternary centres. Vital to our overall synthetic plan was that it required no functionality-masking by means of protecting or caging groups[19]. Such syntheses are strategically streamlined owing to the lack of protection and deprotection steps, operations that inherently add two chemical manipulations per protecting group to any synthetic plan. Although notable examples of protecting group-free total syntheses exist[20–23], they are difficult to accomplish and remain relatively uncommon, particularly in the catalytic enantioselective total synthesis of complex molecules.

The preparation of the key diketone was initiated by the self-condensation of diallyl succinate **8** using a Claisen–Dieckmann process, which was followed by methylation using potassium carbonate and methyl iodide to form bis(β-ketoester) **7** in 51% yield over two steps (Fig. 2a)[24]. This reaction affords **7** as a 1:1 mixture of racemic ((*S*, *S*)-**7** plus (*R*, *R*)-**7**) and meso diastereomers, and can be performed conveniently in 60-g batches of **8**. (Routine synthetic experimental techniques were used for the preparation, isolation, and analysis of all new compounds and are described in the Supplementary Information.) Having obtained bis(β-ketoester) **7**, we were prepared to expose it to the conditions of our Pd-catalysed enantioselective allylation. Typically, use of such a stereoisomeric mixture in an asymmetric transformation would be deleterious to a total synthesis, because the presence of two pre-existing stereocentres in the substrate could interfere with inherent catalyst selectivity and afford reduced quantities of the desired product[25]. The potential for developing mismatched catalyst–substrate interactions that negatively impact yield and selectivity was a major concern, as was the possibility of a kinetic resolution[26].

The number of possible stereochemical outcomes and the number of pathways leading to each potential product renders the situation quite complex (Fig. 2b). Beginning from the diastereomeric mixture of **7**, the substrate must initially undergo decarboxylative deallylation

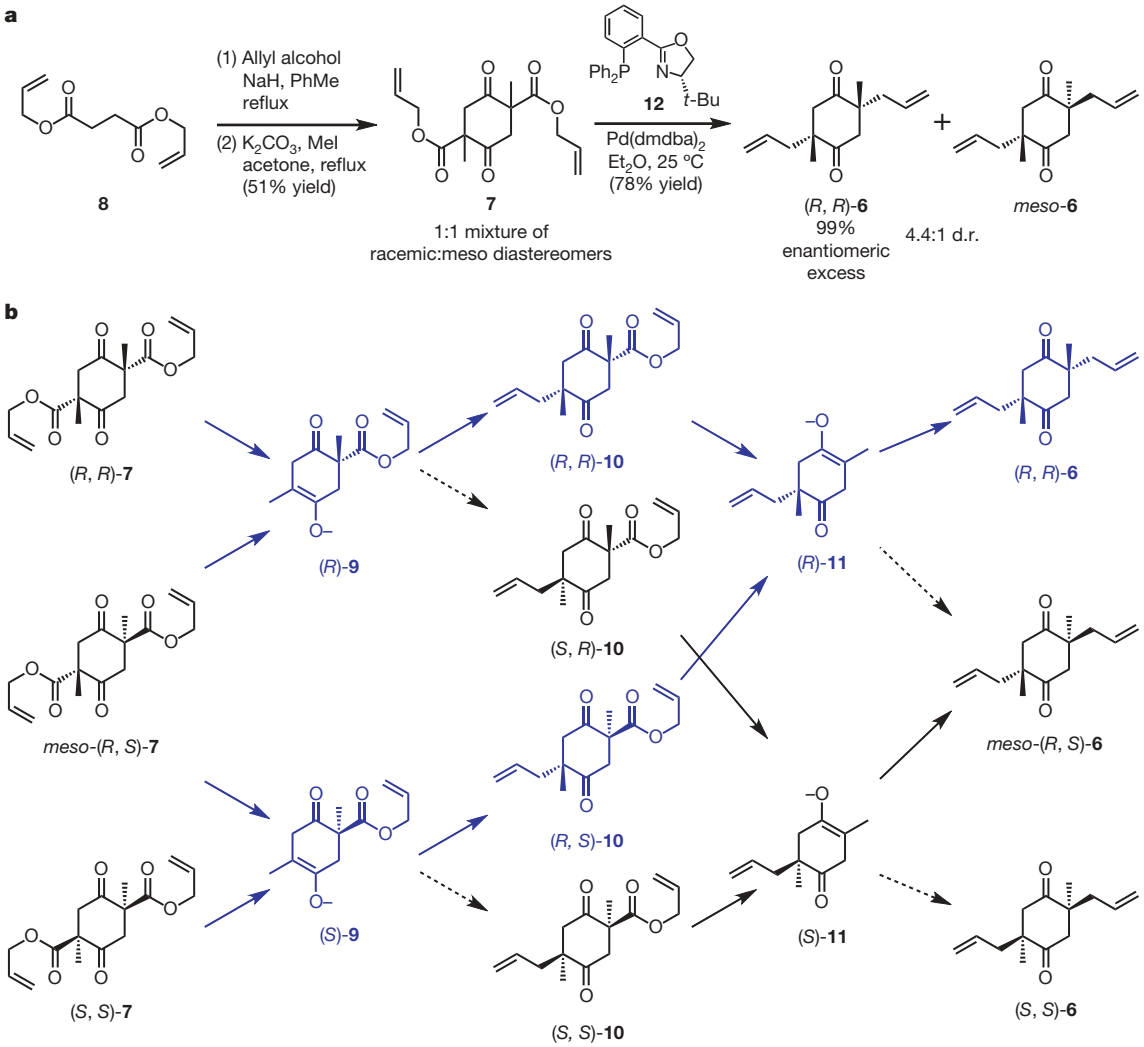


Figure 2 | Synthesis of diketone 6. a, Implementation of the double asymmetric alkylation. **8**→**7**: (1) NaH (2.5 equiv.), allyl alcohol (0.28 equiv.), diallyl succinate (**8**, 1.0 equiv.), PhMe (toluene), 95 °C, 2 h; (2) K_2CO_3 (4.1 equiv.), diallyl succinylsuccinate (1.0 equiv.), MeI (5.1 equiv.), acetone, 50 °C, 6 h, 51% yield over two steps. **7**→**6**: Pd(dmdba)$_2$ (0.05 equiv.), *S*-*t*-BuPHOX (**12**, 0.055 equiv.), Et_2O, 25 °C, precomplexation for 30 min, then bis(β-ketoester) **7** (1.0 equiv.), 25 °C for an additional 10 h, 78% yield, 4.4:1 diastereomeric ratio (d.r.), 99% enantiomeric excess. **b**, Stereochemical analysis for the stereoconvergent double decarboxylative alkylation of **7**. dmdba, bis(3,5-dimethoxybenzylidene)acetone; *t*-BuPHOX, *t*-butyl phosphinooxazoline. For tabulated spectral data of all depicted compounds, please see the Supplementary Information.

to produce racemic enolate **9**, which is subsequently monoalkylated to form ester **10**. If the enantiopure catalyst controls formation of the new stereocentre[17,18], a mixture of ester diastereomers **10** will arise, with (*R*, *R*)-**10** and (*R*, *S*)-**10** each predominating over both (*S*, *R*)-**10** and (*S*, *S*)-**10**. Influence of the remaining substrate stereocentre may either reinforce or conflict with catalyst control during this process. From these intermediate monoesters, a second stereoablative process would then generate a non-racemic mixture of enolates (*R*)-**11** and (*S*)-**11**. Last, a second facially selective alkylation affords the double alkylation products as a diastereomeric mixture of enantioenriched (*R*, *R*)-**6** and *meso*-**6**. For the reaction to proceed favourably a number of criteria must be met: the catalyst must not stereoselectively influence the stereoablative steps appreciably; the catalyst must impose a high degree of facial selectivity at all stages of bond construction; and the diastereoselective bias for alkylation inherent in the substrate must not override the influence of the catalyst. If these requirements are met, it is expected that the major product of the reaction will benefit from a statistical amplification of its enantiomeric excess, in line with theories first described in ref. 3 and later detailed both experimentally and computationally in refs 4–6.

In practice, treatment of a 1:1 mixture of diastereomers (or either pure diastereomer) of bis(β-ketoester) **7** with $Pd(dmdba)_2$ (5 mol per cent) and enantiopure tert-butyl phosphinooxazoline (PHOX) ligand[27] **12** (5.5 mol per cent) in Et_2O at 25 °C affords the bisalkylated products enantioenriched **6** and *meso*-**6** as a 4.4:1 mixture in 78% yield. Moreover, we found that the major diastereomer of **6** is formed in 99% enantiomeric excess. In this single-step procedure, two all-carbon quaternary stereocentres were simultaneously constructed with excellent enantioselectivity, thus addressing what is arguably the greatest synthetic challenge to the preparation of cyanthiwigin F. It is important to note that although this reaction may progress through any of sixteen different pathways to afford any of three different stereoisomers, the high degree of stereoconvergence and stereochemical control imparted by the catalyst system predominantly favours formation of the desired diketone (*R*, *R*)-**6**.

The completion of the synthesis of cyanthiwigin F (**1**) proceeds rapidly owing to the high level of stereochemical value and functionality present in cyclohexanedione **6**. Selective monoenolization of diketone **6** with potassium bis(trimethylsilyl)amide (KHMDS) and trapping of the resulting potassium enolate as a trifluoromethanesulfonate is followed by a Pd-catalysed Negishi cross-coupling to introduce an olefinic side chain and generate tetraene **4** (Fig. 3a)[28]. To advance this material, we required closure of the seven-member ring, as well as elaboration of the terminal allyl group to a functional moiety appropriate for completion of the tricyclic cyathin core. We found that both transformations could be effected with a single catalytic operation by treating **4** with Grubbs' ruthenium catalyst **13** (ref. 29) and a vinyl boronate species (**14**). This provided an efficient and mild method to simultaneously execute both ring-closing metathesis to form the C ring and cross-metathesis with vinyl boronate **14** to elaborate the terminal olefin. Upon oxidative work-up, bicyclic aldehyde **15** was isolated in a single step.

Completion of the carbocyclic core of cyanthiwigin F was then achieved through radical-induced intramolecular cyclization of the aldehyde moiety in bicyclic compound **15** onto the trisubstituted olefin of the central B ring. This was accomplished by treatment of **15** with *t*-BuSH and azobis(isobutylnitrile) (AIBN) at 80 °C to produce tricyclic diketone **16** as a single diastereomer[30]. We found that this reaction afforded only the desired cis-fused A–B ring juncture with an accompanying trans relationship across the B–C ring fusion. We postulate that these configurations result from the kinetic formation of the C3–C4 bond, followed by subsequent kinetic hydrogen atom abstraction at C5 to establish the more thermodynamically stable ring fusion. Recrystallization of this material from acetonitrile and water allowed for unambiguous assignment of the relative stereochemistry by X-ray crystallography (Fig. 3b). Once we had tricyclic compound **16**, selective enol triflate formation was followed by a difficult Pd-catalysed coupling reaction with an *i*-Pr-organocuprate reagent to afford cyanthiwigin F (**1**) and a reduction product (that is, **17**) as a 1.8:1 mixture in 63% combined yield. Isolation of the major component of this mixture affords cyanthiwigin F in only nine steps from diallyl succinate.

In conclusion, we have achieved the catalytic enantioselective synthesis of the cyathin diterpenoid (−)-cyanthiwigin F. Our synthetic route features an enantioselective Pd-catalysed double alkylation reaction as the pivotal step by which stereochemistry is established

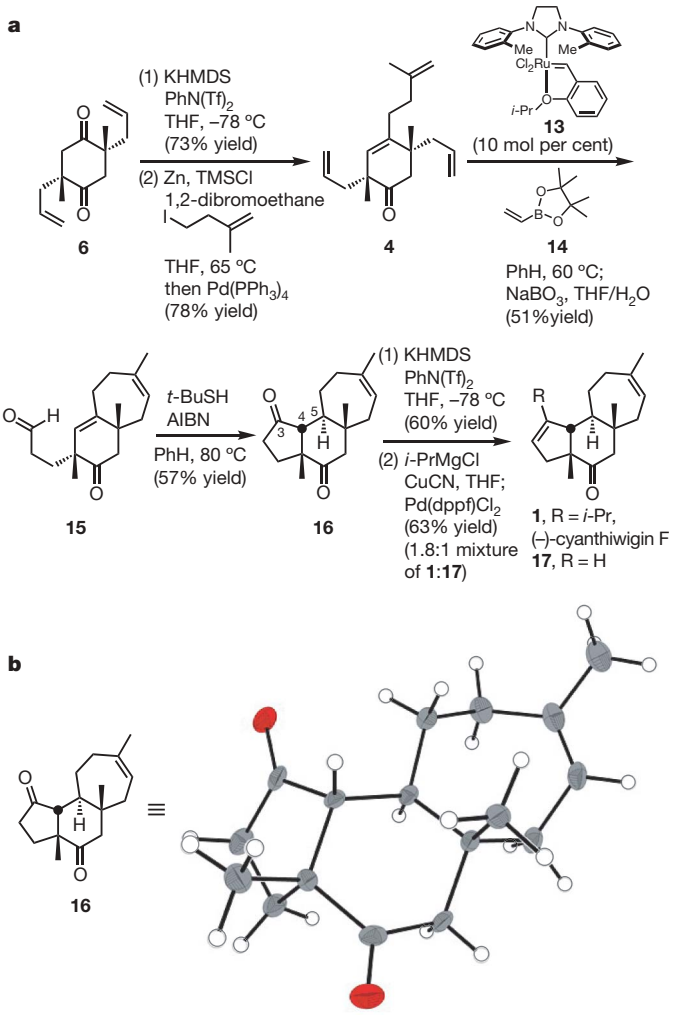


Figure 3 | Synthesis of cyanthiwigin F. a, Completion of the cyanthiwigin F (**1**) synthesis. **6**→**4**: (1) KHMDS (1.1 equiv.), diketone **6** (1.0 equiv.), THF, −78 °C, 30 min, then $PhN(Tf)_2$ (1.2 equiv.), THF, −78 °C, 6 h, 73% yield; (2) Zn powder (7.5 equiv.), 1,2-dibromoethane (1.2 equiv.), TMSCl (0.33 equiv.), THF, 65 °C for 15 min, then 4-iodo-2-methyl-1-butene (1.5 equiv.), 65 °C, 2 h, then enol triflate (1.0 equiv.), $Pd(PPh_3)_4$ (0.05 equiv.), 65 °C, 3 h, 78% yield. **4**→**15**: Tetraene **4** (1.0 equiv.), ruthenium catalyst **13** (0.1 equiv.), benzene, 40 °C, 30 min, then vinyl boronate **14** (5.0 equiv.), 40 °C, 20 h, then $NaBO_3 \cdot H_2O$ (6.0 equiv.), THF/H_2O, 23 °C, 1 h, 51% yield. **15**→**16**: Bicyclic aldehyde **15** (1 equiv.), *t*-butyl thiol (*t*-BuSH, 3.0 equiv.), AIBN (1.5 equiv.), benzene, 80 °C, 22 h, 57% yield. **16**→**1**: (1) KHMDS (1.1 equiv.), diketone **16** (1 equiv.), THF, −78 °C, 30 min, then $PhN(Tf)_2$ (1.15 equiv.), THF, −78 °C, 3 h, 60% yield; (2) CuCN (1.5 equiv.), *i*-PrMgCl (3.0 equiv.), THF, −78 °C to 0 °C, then 0 °C for 10 min, $Pd(dppf)Cl_2$ (0.15 equiv.), enol triflate (1.0 equiv.), THF, 0 °C, 3 h, 63% yield of a 1.8:1 mixture of **1**:**17**. **b**, Oak Ridge Thermal Ellipsoid Plot drawing of **16** (shown with 50% probability ellipsoids). KHMDS, potassium bis(trimethylsilyl)amide; THF, tetrahydrofuran; $PhN(Tf)_2$, phenyl bis(trifluoromethane)sulfonimide; TMSCl, trimethylsilyl chloride; AIBN, 2,2′-azobis(isobutyronitrile); dppf, 1,1′-bis(diphenylphosphino)ferrocene. For tabulated spectral data of all depicted compounds, and crystallographic data of tricyclic diketone **16**, please see the Supplementary Information.

about the central B ring of the molecule. Use of this stereoconvergent approach allowed us to employ classic scaleable condensation chemistry in the early stages of our route (that is, Claisen–Dieckmann cyclization), and provided the critical stereochemical information necessary to guide the final steps of our synthesis. From this point, an efficient tandem ring-closing metathesis/cross-metathesis reaction followed by a thiol-mediated radical cyclization allowed for rapid elaboration to the natural product, without the use of protecting groups. This short synthetic route highlights the utility of double enantioselective transformations in the construction of molecules with high levels of stereochemical complexity, and uses low catalyst loading, is not very difficult operationally, and results in exceptional enantioselectivity. We hope that the results reported here will encourage greater investigation into multiple catalytic enantioselective processes. Implementation of this strategy for the synthesis of other cyanthiwigin natural products and studies towards the use of double enantioselective reactions in the easy preparation of various complex molecules are ongoing.

Received 26 November 2007; accepted 2 May 2008.

1. Masamune, S., Choy, W., Petersen, J. & Sita, L. Double asymmetric synthesis and a new strategy for stereocontrol in organic synthesis. *Angew. Chem. Int. Edn Engl.* **24,** 1–30 (1985).
2. Kolodiazhnyi, O. I. Multiple stereoselectivity and its applications in organic synthesis. *Tetrahedron* **59,** 5953–6018 (2003).
3. Langenbeck, W. & Triem, G. Zur Theorie der Erhaltung und Entstehung optischer Aktivität in der Natur. *Z. Phys. Chem. A* **117,** 401–409 (1936).
4. Vigneron, J. P., Dhaenens, M. & Horeau, A. Nouvelle méthode pour porter au maximum la pureté optique d'un produit partiellement dédoublé sans l'aide d'aucune substance chirale. *Tetrahedron* **29,** 1055–1059 (1973).
5. Rautenstrauch, V. The two expressions of the Horeau principle, nth-order Horeau amplifications, and scales for the resulting very high enantiopurities. *Bull. Soc. Chim. Fr.* **131,** 515–524 (1994).
6. Baba, S. E., Sartor, K., Poulin, J. & Kagan, H. Tandem asymmetric syntheses from achiral precursors – asymmetric homogeneous reduction of bisdehydrodipeptides. *Bull. Soc. Chim. Fr.* **131,** 525–533 (1994).
7. Mohr, J. T., Ebner, D. C. & Stoltz, B. M. Catalytic enantioselective stereoablative reactions: an unexploited approach to enantioselective catalysis. *Org. Biomol. Chem.* **5,** 3571–3576 (2007).
8. Peng, J. *et al.* The new bioactive diterpenes cyanthiwigin E-AA from the Jamaican sponge *Myrmekioderma styx. Tetrahedron* **58,** 7809–7819 (2002).
9. Sennett, S. H., Pomponi, S. A. & Wright, A. E. Diterpene metabolites from two chemotypes of the marine sponge *Myrmekioderma styx. J. Nat. Prod.* **55,** 1421–1429 (1992).
10. Peng, J., Avery, M. A. & Hamann, M. T. Cyanthiwigin AC and AD, two novel diterpene skeletons from the Jamaican sponge *Myrmekioderma styx. Org. Lett.* **5,** 4575–4578 (2003).
11. Saito, T. *et al.* Erinacine E as a kappa opioid receptor agonist and its new analogs from a basidiomycete, *Hericium ramosum*. *J. Antibiot. (Tokyo)* **51,** 983–990 (1998).
12. Cozzi, P. G., Hlgraf, R. & Zimmermann, N. Enantioselective catalytic formation of quaternary stereogenic centers. *Eur. J. Org. Chem.* **2007,** 5969–5994 (2007).
13. Trost, B. M. & Jiang, C. Catalytic enantioselective construction of all-carbon quaternary stereocenters. *Synthesis* 369–396 (2006).
14. Pfeiffer, M. W. B. & Phillips, A. J. Total synthesis of (+)-cyanthiwigin U. *J. Am. Chem. Soc.* **127,** 5334–5335 (2005).
15. Reddy, J. T., Bordeau, G. & Trimble, L. Total synthesis of (+)-cyanthiwigin AC. *Org. Lett.* **8,** 5585–5588 (2006).
16. Poss, C. S. & Schreiber, S. L. Two-directional chain synthesis and terminus differentiation. *Acc. Chem. Res.* **27,** 9–17 (1994).
17. Behenna, D. C. & Stoltz, B. M. The enantioselective Tsuji allylation. *J. Am. Chem. Soc.* **126,** 15044–15045 (2004).
18. Mohr, J. T., Behenna, D. C., Harned, A. M. & Stoltz, B. M. Deracemization of quaternary stereocenters by Pd-catalyzed enantioconvergent decarboxylative allylation of racemic β-ketoesters. *Angew. Chem. Int. Edn Engl.* **44,** 6924–6927 (2005).
19. Greene, T. & Wuts, P. *Protective Groups in Organic Synthesis*. (Wiley, New York, 1999).
20. Robinson, R. A synthesis of tropinone. *J. Chem. Soc.* **111,** 762–768 (1917).
21. Hoffmann, R. W. Protecting-group-free synthesis. *Synthesis* 3531–3541 (2006).
22. McFadden, R. M. & Stoltz, B. M. The catalytic enantioselective, protecting group-free total synthesis of (+)-dichroanone. *J. Am. Chem. Soc.* **128,** 7738–7739 (2006).
23. Baran, P. S., Maimone, T. J. & Richter, J. M. Total synthesis of marine natural products without using protecting groups. *Nature* **446,** 404–408 (2007).
24. Ebert, H. Zur Constitution des Succinylobernsteinsäureäthers. *Liebigs Ann. Chem.* **229,** 45–88 (1885).
25. Kagan, H. Various aspects of the reaction of a chiral catalyst or reagent with a racemic or enantiopure substrate. *Tetrahedron* **57,** 2449–2468 (2001).
26. Eliel, E. L. & Wilen, S. H. *Stereochemistry of Organic Compounds* 965–971 (Wiley, New York, 1994).
27. Helmchen, G. & Pfaltz, A. Phosphinooxazolines-a new class of versatile, modular P,N-ligands for asymmetric catalysis. *Acc. Chem. Res.* **33,** 336–345 (2000).
28. Taishi, T., Takechi, S. & Mori, S. First total synthesis of (±)-stachyflin. *Tetrahedr. Lett.* **39,** 4347–4350 (1998).
29. Stewart, I. C. *et al.* Highly efficient ruthenium catalysts for the formation of tetrasubstituted olefins via ring-closing metathesis. *Org. Lett.* **9,** 1589–1592 (2007).
30. Yoshikai, K., Hayama, T., Nishimura, K., Yamada, K. & Tomioka, K. Thiol-catalyzed acyl radical cyclization of alkenals. *J. Org. Chem.* **70,** 681–683 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors wish to thank NIH-NIGMS (R01GM080269-01), Amgen, Abbott, Boehringer Ingelheim, Merck and Bristol-Myers Squibb for financial support. We also wish to thank M. W. Day and L. M. Henling for X-ray crystallographic expertise, S. Virgil, A. Harned, D. White, D. Caspi and J. T. Mohr for helpful discussions, and M. T. Hamann for an authentic sample and spectra of cyanthiwigin F. We thank E. J. Corey for guidance and mentorship, on the occasion of his 80th birthday.

Author Information Crystallographic data have been deposited at the Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK, and copies can be obtained on request, free of charge, by quoting the publication citation and the deposition number 664430. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to B.M.S. (stoltz@caltech.edu).

LETTERS

Extensive halogen-mediated ozone destruction over the tropical Atlantic Ocean

Katie A. Read[1], Anoop S. Mahajan[2], Lucy J. Carpenter[1], Mathew J. Evans[3], Bruno V. E. Faria[4], Dwayne E. Heard[2], James R. Hopkins[5], James D. Lee[5], Sarah J. Moller[1], Alastair C. Lewis[5], Luis Mendes[4], James B. McQuaid[3], Hilke Oetjen[2], Alfonso Saiz-Lopez[6], Michael J. Pilling[2] & John M. C. Plane[2]

Increasing tropospheric ozone levels over the past 150 years have led to a significant climate perturbation[1]; the prediction of future trends in tropospheric ozone will require a full understanding of both its precursor emissions and its destruction processes. A large proportion of tropospheric ozone loss occurs in the tropical marine boundary layer[2,3] and is thought to be driven primarily by high ozone photolysis rates in the presence of high concentrations of water vapour. A further reduction in the tropospheric ozone burden through bromine and iodine emitted from open-ocean marine sources has been postulated by numerical models[4–7], but thus far has not been verified by observations. Here we report eight months of spectroscopic measurements at the Cape Verde Observatory indicative of the ubiquitous daytime presence of bromine monoxide and iodine monoxide in the tropical marine boundary layer. A year-round data set of co-located *in situ* surface trace gas measurements made in conjunction with low-level aircraft observations shows that the mean daily observed ozone loss is ~50 per cent greater than that simulated by a global chemistry model using a classical photochemistry scheme that excludes halogen chemistry. We perform box model calculations that indicate that the observed halogen concentrations induce the extra ozone loss required for the models to match observations. Our results show that halogen chemistry has a significant and extensive influence on photochemical ozone loss in the tropical Atlantic Ocean boundary layer. The omission of halogen sources and their chemistry in atmospheric models may lead to significant errors in calculations of global ozone budgets, tropospheric oxidizing capacity and methane oxidation rates, both historically and in the future.

Tropospheric ozone is an important greenhouse gas in addition to its influence on air quality and public health, on the photochemical processing of atmospheric chemicals, and on food security and ecosystem viability. It is produced through the catalytic oxidation of carbon compounds in the presence of nitrogen oxides ($NO_x = NO + NO_2$), and has an additional smaller source from stratospheric influx of ozone into the free troposphere[8]. Ozone is lost to the surface through deposition and can be destroyed throughout the atmosphere by photochemical processes, predominantly by photolysis and the subsequent reaction of electronically excited oxygen atoms with water vapour. Whether a particular air mass is producing or destroying ozone depends broadly on the short-wave radiation environment and the water vapour and NO_x concentrations. Thus, ozone is formed predominantly in continental regions where there are sources of NO_x and is typically lost in marine regions where sources are small[9]. Because of its high water vapour content, high solar radiation levels and large geographical extent, the tropical marine boundary layer is the most important global region for loss of ozone[10]. However, surface atmospheric observations in this region are extremely sparse.

The Cape Verde archipelago lies within the tropical Eastern North Atlantic Ocean. The archipelago is volcanic in origin and the island shores shelve steeply to the deep abyssal plain beyond the African continental shelf. An Ocean–Atmosphere Observatory has been newly established on the northeastern side of the island of São Vicente within the Cape Verde archipelago. The atmospheric site (16.85° N, 24.87° W, Supplementary Fig. 1) receives the prevailing northeasterly trade winds directly off the ocean for around 95% of the time. In contrast to many other atmospheric monitoring stations in the Northern Hemisphere, there are no seaweed beds or other local coastal sources. It can therefore be assumed to be representative of the surrounding open-ocean marine boundary layer. The ocean surrounding Cape Verde is in general biologically productive because of both Saharan dust input[11] and proximity to the northwest African coastal upwelling system, which lies a few hundred kilometres to the northeast.

A year of observations from the Observatory has now been obtained (October 2006 to October 2007). These include measurements of O_3, H_2O, CO, NO, NO_2, CH_4, volatile organic compounds (VOCs), and oxygenated VOCs, dimethyl sulphide, the halogen oxide radicals BrO, IO and OIO, the ozone photolysis rate coefficient $J(O^1D)$, broadband ultraviolet radiation, wind speed and direction (see Supplementary Information). During the summer of 2007, a research aircraft measured composition over the site to assess the representativeness of the overlying boundary layer and to determine any diurnal variability in boundary layer depth (see Supplementary Information).

The annual cycle in ozone (see Supplementary Fig. 2) displays a maximum in spring and a minimum in late summer, consistent with other coastal sites[12,13]. The daily cycle shows a loss during the day and a recovery at night (due to entrainment from the free troposphere) with an annually averaged loss of 3.3 ± 2.6 p.p.b.v. (parts per 10^9 by volume) per day between 09:00 and 17:00 h UT (local plus 1 h). The aircraft observations in June 2007 confirm the vertical extent of the ozone loss throughout the boundary layer (Fig. 1). This agreement between surface and aircraft-determined ozone concentration in the boundary layer is typical and was observed on 12 further flights not shown here.

The monthly averaged daytime (09:00–17:00 UT) concentrations of ozone and NO observed at Cape Verde between October 2006 and

[1]Department of Chemistry, University of York, Heslington, York YO10 5DD, UK. [2]School of Chemistry, University of Leeds, Leeds, LS2 9JT, UK. [3]School of Earth and the Environment (SEE), University of Leeds, LS2 9JT, UK. [4]Instituto Nacional de Meteorologia Geofísica (INMG), Delegação de São Vicente, Monte, CP 15, Mindelo, Cape Verde. [5]National Centre for Atmospheric Science (NCAS), University of York, Heslington, York YO10 5DD, UK. [6]Earth and Space Science Division, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA.

October 2007 were compared with simulations using the global tropospheric chemistry transport model GEOS-CHEM[14]. The measured NO mixing ratios were extremely low throughout the year (09:00–17:00 UT average 3.0 ± 1.0 p.p.t.v. (1σ) where p.p.t.v. is parts per 10^{12} by volume) and showed broad agreement with the GEOS-CHEM simulated levels (09:00–17:00 UT average 3.2 ± 1.3 p.p.t.v.; Fig. 2). However, the ozone observations show a significantly greater monthly averaged daytime depletion of ozone (the difference between concentrations at 09:00 and 17:00 UT, referred to hereafter as ΔO_3) and a more exaggerated seasonal cycle in ΔO_3 than the model simulations. The ozone budget in GEOS-CHEM includes NO_x-catalysed ozone production and odd-hydrogen photochemical ozone loss, in addition to advection, convection and deposition terms. During the period of maximum solar activity, GEOS-CHEM computes a maximum net ozone loss of ~3 p.p.b.v. d^{-1} (up to 8% of the mean GEOS-CHEM ozone concentration) compared with the observed losses of ~5 p.p.b.v. d^{-1} (up to 13% of the mean measured ozone concentration) (Fig. 2). Similar results are obtained from box model simulations (see Supplementary Information) of the O_x–HO_x–NO_x system, constrained using the observations.

Because of slowly changing sea surface temperatures and continuous strong winds, the marine boundary layer around Cape Verde and the Observatory is subject to minimal diurnal variations in dynamics such as boundary layer depth, which can complicate interpretations and modelling of changes in O_3 concentrations. The surface temperature varies by no more than ±0.5 °C throughout the day and drops by only 1 °C at night. Aircraft measurements made on four days, each day with flights typically at 09:00, 13:00 and 18:30 UT, confirm negligible changes in the inversion depth over any given day (Fig. 1). Other key parameters for accurately simulating the daily ozone variation in low NO_x regions are the entrainment rate—the largest ozone budget term in this region—and the photolytic destruction rate. For the box model simulations, the observed average monthly nocturnal increase in ozone at Cape Verde was assumed to comprise the difference between the entrainment and deposition terms over the month. This parameter demonstrated a seasonal variation with a minimum in November (0.18 p.p.b.v. h^{-1}) and a maximum in April (0.48 p.p.b.v. h^{-1}). Hourly GEOS-CHEM values were used for the ozone photolysis rate coefficient $J(O^1D)$. These are calculated using Fast-J code[15] which uses the cloud fields from the Goddard Earth Observing System of the NASA Global Modelling and Assimilation Office (GMAO), and ozone columns derived from satellite observations. The calculated monthly averaged peak values of $J(O^1D)$ agreed with data obtained in January–February and May–June 2007 to within 13%.

That both modelling approaches significantly underestimate the observed daily ozone loss is clearly evidence for an important missing loss process. A widespread effect of halogens on tropospheric oxidants in the marine boundary layer has been suggested by a number of theoretical[4,5,7] and observational[16,17] studies; however, these so far remain unconfirmed predominantly because of a lack of observations of atmospheric species representative of the pristine marine boundary layer[18,19]. Ozone is destroyed directly via catalytic cycles with the rate determined by the self-reactions of XO (where XO = IO, BrO) and the reaction of XO with HO_2 (reactions (1)–(5)). A reduction in ozone production also occurs through a decrease of the HO_2/OH

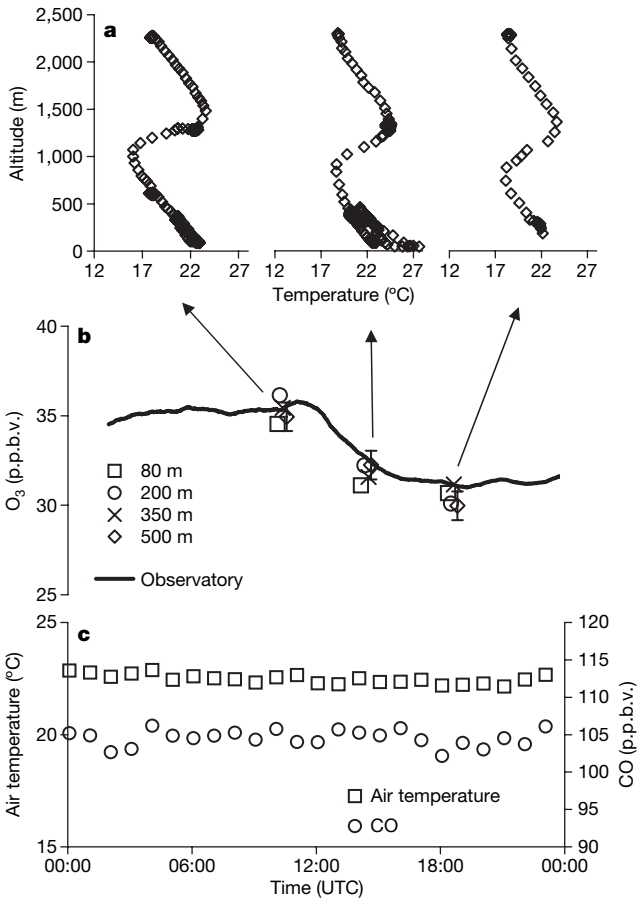


Figure 1 | Example data from aircraft observations 1–20 km upwind of São Vicente. **a**, Vertical temperature profiles obtained at 09:00, 13:00 and 18:30 UT on 27 May 2007 showing the constant inversion height. **b**, Mean ozone measured at four vertical levels from the aircraft (symbols, 10-min averages) and at the Cape Verde Observatory (solid line, 20-min running mean) on 27 May 2007. The 1σ standard deviations of the mean of the aircraft ozone data from 200 m are shown as an example of the precision of the data (see Supplementary Information). c, Temperature and CO on 27 May 2007.

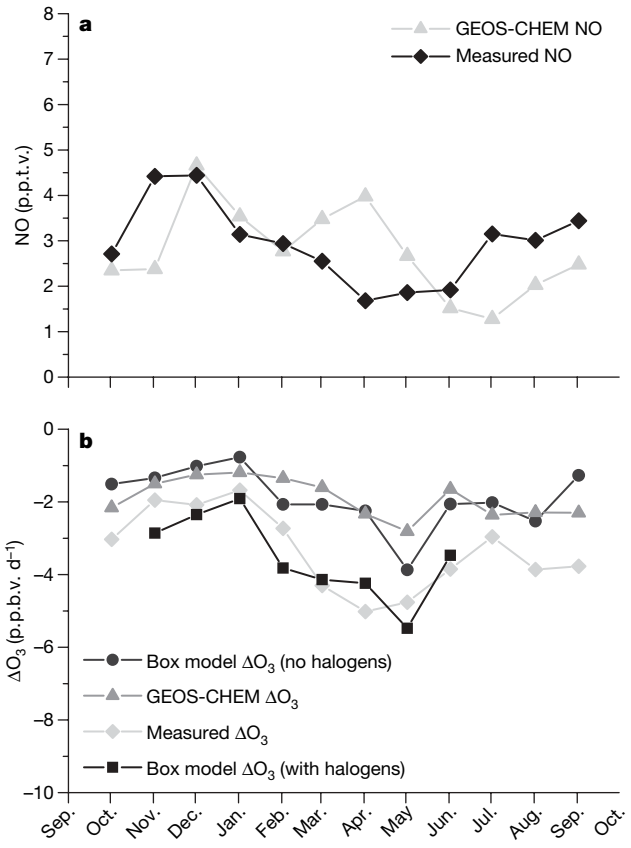


Figure 2 | Measurements and modelling results. **a**, Monthly averages of observed and modelled (GEOS-CHEM) NO mixing ratios; **b**, monthly averages of observed ΔO_3 over 8 h (09:00–17:00 UT) compared with predictions from GEOS-CHEM and from the box model with and without halogen chemistry.

ratio as a result of reactions (4) and (5), and the suppression of NO_x caused by the hydrolysis of halogen nitrates on aerosol surfaces[20].

$$X + O_3 \rightarrow XO + O_2 \quad (1)$$

$$XO + XO \rightarrow 2X + O_2 \quad (2)$$

$$XO + IO \rightarrow X + OIO \quad (3)$$

$$XO + HO_2 \rightarrow HOX + O_2 \quad (4)$$

$$HOX + h\nu \rightarrow OH + X \quad (5)$$

$$XO + NO \rightarrow NO_2 + X \quad (6)$$

Although halogen oxides also cause a decrease in the NO/NO_2 ratio through reaction (6), the subsequent formation of NO_2 does not necessarily lead to an increase in ozone concentrations because a halogen atom is also formed which will destroy ozone through reaction (1). In addition to its effect on ozone and OH, it has been proposed that BrO causes marked changes in dimethyl sulphide levels and oxidation pathways, reducing its cooling effect on climate[5,21].

Halogen oxide measurements by differential optical absorption spectroscopy (Supplementary Information) at the Cape Verde Observatory from November 2006 until June 2007 reveal the presence in daytime of IO and BrO radicals with mean daytime maxima of 1.4 ± 0.8 (1σ) p.p.t.v. and 2.5 ± 1.1 (1σ) p.p.t.v., respectively (Fig. 3). Both radicals show diurnal cycles that seem to be dependent on solar radiation, with mixing ratios below the detection limits (BrO: 0.5–1 p.p.t.v., IO: 0.3–0.5 p.p.t.v.) at night. On average, slightly higher average values are apparent in spring; but the variability of the observations does not allow for a robust conclusion on seasonal halogen oxide trends. Models for the acid-catalysed activation of bromine from sea salt aerosol predict a marine boundary layer BrO concentration of about 1–4 p.p.t.v. with a diurnal cycle similar to a top-hat distribution[4,22]. Iodine oxide levels of about 1 p.p.t.v. with a similar diurnal profile were predicted from photolysis of 'typical' quantities of organoiodine compounds in the marine boundary layer[4], although there are large uncertainties in such predictions, especially because of the lack of data on the iodine atom precursors. The concentrations and diurnal behaviour of IO and BrO are thus consistent with prior understanding.

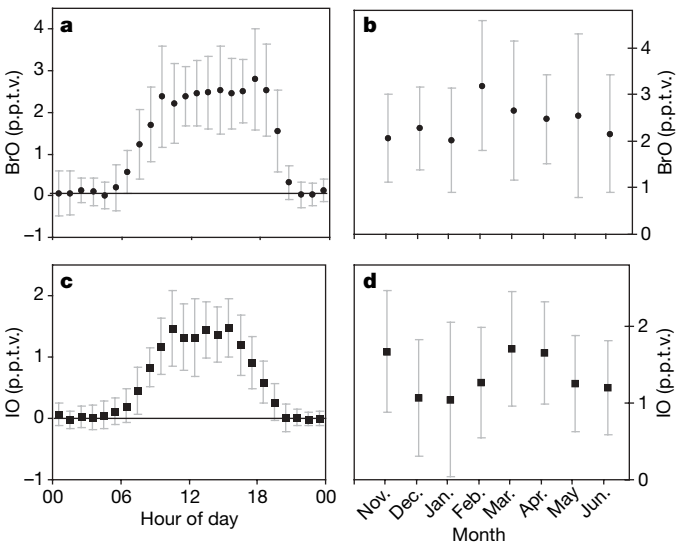


Figure 3 | Halogen oxide observations. Averaged diurnal profiles for BrO (**a**) and IO (**c**). Errors (1σ) are indicated as grey lines. **b**, Seasonal variation in BrO; **d**, seasonal variation in IO. The points show average concentrations seen from 09:00 to 17:00 UT.

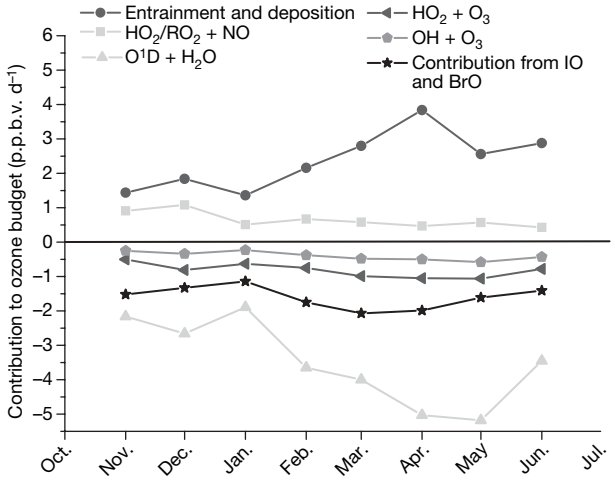


Figure 4 | Monthly averaged contributions to the daily ozone budgets between 09:00 and 17:00 UT. The contribution from halogen oxides, calculated using the measured concentrations from Nov 2006 to Jun 2007, takes into account both the direct and indirect effects on ozone (see text).

When observed hourly averaged halogen concentrations are included in the box model, the simulated annual average daily ozone loss (3.2 ± 1.1 (1σ) p.p.b.v. d^{-1} between 09:00 and 17:00 UT) is comparable to that observed, and the monthly average cycle in ozone loss is better reproduced (Fig. 2). The ozone loss is greatest between March and May and is thought to be a consequence of increased photolysis rates and therefore increased photochemical ozone loss in addition to seasonally changing halogen oxide concentrations. The halogen-mediated O_3 destruction contributes an average of 1.8 ± 0.4 p.p.b.v. d^{-1} to the total ozone loss (Fig. 4). Contributions of 0.35, 1.24 and 2.07 p.p.b.v. d^{-1} are attributed to BrO, IO and the sum of IO and BrO, respectively, during the maximum in March. The effect of IO and BrO together is greater than the sum of their individual contributions because of the IO + BrO reaction (reaction (3)). If halogens are excluded in the box model calculations, then the model underestimates the O_3 loss by 47% and overestimates the O_3 concentration by 12% (annual averages). Exclusion of halogens also leads to an underestimation of modelled OH concentrations by 5–12% in this region when compared with the full model concentrations, with corresponding overestimates of the lifetimes of important trace gases such as methane. A compelling argument for exclusion of localized sources of halogens as an explanation for the observations reported here is that if the halogen sources were restricted to the kilometre around the island, the ozone loss rate required to explain the measurements would be more than 100 p.p.b.v. h^{-1}, requiring a halogen loading that is orders of magnitude higher than any tropospheric observations.

The combination of long-term localized chemical data with geographically widespread boundary layer O_3 loss seen from aircraft leads us to the conclusion that halogens play an important role in this region and that their chemical influence extends at least over several thousand kilometres. Current understanding of the sources of bromine indicates that there is no a priori reason that this region should not be representative of bromine levels in the oceanic marine boundary layer in general[6]. The oceans around Cape Verde are biologically active, however, and it is possible that they represent an area of increased iodine release. The inclusion of halogen sources and chemistry into global climate models is now essential for understanding ozone as a climate gas, and for calculating tropospheric oxidizing capacity and the methane lifetime.

METHODS SUMMARY

Ozone was measured every minute from a height of 5 m using an ultraviolet absorption instrument (Model 49i Thermo Scientific). The instrument gives an

absolute measurement of ozone with a precision of 0.07 p.p.b.v. for hourly averaged data. Mixing ratios for IO and BrO were retrieved from long-path differential optical absorption spectroscopy[23,24] using a newtonian telescope acting as a transmitter and receiver, and an array of retro-reflectors placed 6.1 km across a bay from the observatory. Spectra were collected every 30 s and then further averaged over 20–30 min to improve the signal-to-noise ratio. Mixing ratios for CO from 10 m were determined using an Aerolaser 5001 fast response VUV analyser, with a detection limit of <0.5 p.p.b.v. on 1-min average. Measurements of NO_x were made from 5 m using a single channel, chemiluminescence NO detector with a photolytic NO_2 converter. The instrument[25] (Air Quality Design) alternates between measuring NO and NO_2 in a 10-min duty cycle, with hourly averaged data giving detection limits of 1.5 p.p.t.v. and 4 p.p.t.v. for NO and NO_2, respectively. Hourly VOC measurements (C_2–C_8 non-methane hydrocarbons, dimethyl sulphide and C_1–C_3 oxygenated VOC) were made from 10 m using a dual-channel gas chromatograph with flame ionization detection[26]. Detection limits for VOC ranged between 5 and 10 p.p.t.v. Weekly CH_4 measurements were obtained using flask samples with subsequent gas chromatographic analysis. Temperature, relative humidity and wind measurements were collected from 10 m at 1 Hz and then averaged over one minute. Atmospheric pressure and broadband ultraviolet radiation were recorded at 4 m. The rate coefficient $J(O^1D)$ was measured at 4 m with a 2π filter radiometer (Meteorology Consult) with $<5\%$ precision and $\sim 20\%$ accuracy for solar zenith angles below 60°, and corrections were applied for the vertical overhead O_3 column and changes in sensitivity due to changing solar zenith angle.

Received 28 January; accepted 21 April 2008.

1. Intergovernmental Panel on Climate Change (IPCC) *Climate Change 2007: The Physical Sciences Basis*, available at ⟨http://ipcc-wg1.ucar.edu/wg1/wg1-report.html⟩. (Retrieved on 30 April 2007.)
2. Horowitz, L. W. *et al.* A global simulation of tropospheric ozone and related tracers: Description and evaluation of MOZART, version 2. *J. Geophys. Res.* **108** (D24), 4784–4812 (2003).
3. Lawrence, M. G., Jockel, P. & von Kuhlmann, R. What does the global mean OH concentration tell us? *Atmos. Chem. Phys.* **1,** 37–49 (2001).
4. Vogt, R., Sander, R., von Glasow, R. & Crutzen, P. J. Iodine chemistry and its role in halogen activation and ozone loss in the marine boundary layer: A model study. *J. Atmos. Chem.* **32,** 375–395 (1999).
5. von Glasow, R., von Kuhlmann, R., Lawrence, M. G., Platt, U. & Crutzen, P. J. Impact of reactive bromine chemistry in the troposphere. *Atmos. Chem. Phys.* **4,** 2481–2497 (2004).
6. Yang, X. *et al.* Tropospheric bromine chemistry and its impacts on ozone: A model study. *J. Geophys. Res.* **110,** D23311 (2005).
7. von Glasow, R., Sander, R., Bott, A. & Crutzen, P. J. Modelling halogen chemistry in the marine boundary layer. 1. Cloud-free MBL. *J. Geophys. Res.* **107** (D17), 4341–4356 (2002).
8. Junge, C. E. Global ozone budget and exchange between stratosphere and troposphere. *Tellus* **14,** 363–377 (1962).
9. Lelieveld, J. *et al.* Increasing ozone over the Atlantic Ocean. *Science* **304,** 1483–1487 (2004).
10. Bloss, W. J. *et al.* The oxidative capacity of the troposphere: Coupling of field measurements of OH and a global chemistry transport model. *Faraday Discuss.* **130,** 425–436 (2005).
11. Falkowski, P. G. Evolution of the nitrogen cycle and its influence on the biological sequestration of CO_2 in the ocean. *Nature* **387,** 272–274 (1997).
12. Simmonds, P. G., Derwent, R. G., Manning, A. L. & Spain, G. Significant growth in surface ozone at Mace Head, Ireland, 1987–2003. *Atmos. Environ.* **38,** 4769–4778 (2004).
13. Parrish, D. D. *et al.* Relationships between ozone and carbon monoxide at surface sites in the North Atlantic region. *J. Geophys. Res.* **103** (D11), 13357–13376 (1998).
14. Bey, I. *et al.* Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation. *J. Geophys. Res.* **106,** 23073–23095 (2001).
15. Wild, O., Zhu, X. & Prather, M. J. Fast-J: accurate simulation of in- and below-cloud photolysis in tropospheric chemical models. *J. Atmos. Chem.* **37,** 245–282 (2004).
16. Galbally, I. E., Bentley, S. T. & Meyer, C. P. Mid-latitude marine boundary-layer ozone destruction at visible sunrise observed at Cape Grim, Tasmania, 41 degrees S. *Geophys. Res. Lett.* **27,** 3841–3844 (2000).
17. Dickerson, R. R. *et al.* Ozone in the remote marine boundary layer: A possible role for halogens. *J. Geophys. Res.* **104,** 21385–21396 (1999).
18. Allan, B. J., McFiggans, G., Plane, J. M. C. & Coe, H. The nitrate radical in the remote marine boundary layer. *J. Geophys. Res.* **105,** 24191–24204 (2000).
19. Leser, H., Honninger, G. & Platt, U. MAX-DOAS measurements of BrO and NO_2 in the marine boundary layer. *Geophys. Res. Lett.* 30, art. no. 1537 (2003).
20. Sander, R., Rudich, Y., von Glasow, R. & Crutzen, P. J. The role of $BrNO_3$ in marine tropospheric chemistry: A model study. *Geophys. Res. Lett.* **26,** 2857–2860 (1999).
21. Toumi, R. BrO as a sink for dimethylsulfide in the marine atmosphere. *Geophys. Res. Lett.* **21,** 117–120 (1994).
22. Vogt, R., Crutzen, P. J. & Sander, R. A mechanism for halogen release from sea-salt aerosol in the remote marine boundary layer. *Nature* **383,** 327–330 (1996).
23. Plane, J. M. C. & Saiz-Lopez, A. in *Analytical Techniques for Atmospheric Measurement* (ed. Heard, D. E.) (Blackwell, Oxford, 2006).
24. Platt, U. in *Air Monitoring by Spectroscopy Techniques* (ed. Sigrist, M. W.) 27–83 (Wiley, London, 1994).
25. Davis, D. *et al.* South Pole NO_x Chemistry: An assessment of factors controlling variability and absolute levels. *Atmos. Environ.* **38,** 5375–5388 (2004).
26. Hopkins, J. R., Lewis, A. C. & Read, K. A. A two-column method for long-term monitoring of non-methane hydrocarbons (NMHCs) and oxygenated volatile organic compounds (o-VOCs). *J. Environ. Monit.* **4,** 1–7 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank pilots C. Joseph and D. Davies from the NERC Airborne Research and Support Facility, Oxford, for their assistance in obtaining the vertically resolved observations. We thank M. Heimann for provision of CH_4 data from the Cape Verde Observatory and K. Furneaux and L. Whalley for provision of $J(O^1D)$ data. We acknowledge the UK NERC Surface Ocean Lower Atmosphere programme and the EU (Tropical Eastern North Atlantic Time Series Observatory) for funding. Finally, we thank D. Wallace, M. Heimann, J. Pimenta Lima and O. Melicio for their roles in setting up the Cape Verde Observatory, and G. McFiggans for conception of the Reactive Halogens in the Marine Boundary Layer Experiment, which contributed to this paper.

Author Contributions L.J.C, J.M.C.P, M.J.P. and A.C.L. conceived the experiment, and together with K.A.R., A.S.M., B.V.E.F., D.E.H., J.R.H., J.D.L, S.J.M, L.M., J.B.M., H.O. and A.S.-L. carried it out; L.J.C., M.J.E and K.A.R. carried out the data analysis; L.J.C., A.C.L, M.J.E and K.A.R. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to L.J.C. (ljc4@york.ac.uk) or J.M.C.P. (j.m.c.plane@leeds.ac.uk).

LETTERS

Explosive volcanism on the ultraslow-spreading Gakkel ridge, Arctic Ocean

Robert A. Sohn[1], Claire Willis[1], Susan Humphris[1], Timothy M. Shank[1], Hanumant Singh[1], Henrietta N. Edmonds[2], Clayton Kunz[1], Ulf Hedman[3], Elisabeth Helmke[4], Michael Jakuba[5], Bengt Liljebladh[6], Julia Linder[4], Christopher Murphy[1], Ko-ichi Nakamura[7], Taichi Sato[8], Vera Schlindwein[4], Christian Stranne[6], Maria Tausenfreund[4], Lucia Upchurch[2], Peter Winsor[1], Martin Jakobsson[9] & Adam Soule[1]

Roughly 60% of the Earth's outer surface is composed of oceanic crust formed by volcanic processes at mid-ocean ridges. Although only a small fraction of this vast volcanic terrain has been visually surveyed or sampled, the available evidence suggests that explosive eruptions are rare on mid-ocean ridges, particularly at depths below the critical point for seawater (3,000 m)[1]. A pyroclastic deposit has never been observed on the sea floor below 3,000 m, presumably because the volatile content of mid-ocean-ridge basalts is generally too low to produce the gas fractions required for fragmenting a magma at such high hydrostatic pressure. We employed new deep submergence technologies during an International Polar Year expedition to the Gakkel ridge in the Arctic Basin at 85° E, to acquire photographic and video images of 'zero-age' volcanic terrain on this remote, ice-covered ridge. Here we present images revealing that the axial valley at 4,000 m water depth is blanketed with unconsolidated pyroclastic deposits, including bubble wall fragments (limu o Pele)[2], covering a large (>10 km^2) area. At least 13.5 wt% CO_2 is necessary to fragment magma at these depths[3], which is about tenfold the highest values previously measured in a mid-ocean-ridge basalt[4]. These observations raise important questions about the accumulation and discharge of magmatic volatiles at ultraslow spreading rates on the Gakkel ridge[5] and demonstrate that large-scale pyroclastic activity is possible along even the deepest portions of the global mid-ocean ridge volcanic system.

The Gakkel ridge, stretching ~1,800 km across the eastern Arctic Basin, is the ultraslow-spreading end-member of the global mid-ocean ridge (MOR) system, and in 1999 the Global Seismic Network (GSN) detected the largest MOR earthquake swarm ever recorded[6] on the ridge at 85° E. Several lines of evidence suggest that the swarm was associated with a major volcanic event[6–10], but our ability to characterize volcanic processes in this region has been limited by its remote location and ice cover. From 15 to 31 July 2007 the Arctic Gakkel Vents (AGAVE) expedition aboard the icebreaker *Oden* surveyed the presumed eruption site with a Kongsberg EM120 1° × 1° multibeam echo sounder, a conductivity–temperature–depth rosette, two autonomous underwater vehicles and a sub-ice camera and sampling platform (CAMPER).

We produced a high-resolution bathymetric map of the axial valley floor at 85° E by operating the sonar system while drifting quiescently in the ice pack. The combination of the low-noise survey mode and the decrease in variance obtained from ping-averaging several dozen overlapping tracklines allowed us to produce a highly detailed (30-m pixel resolution) sonar image of the eruption site (Fig. 1), showing that the axial valley is filled with distinctive volcanic features. These volcanoes are up to ~2,000 m in diameter and a few hundred metres high. They are commonly flat-topped, contain a prominent central

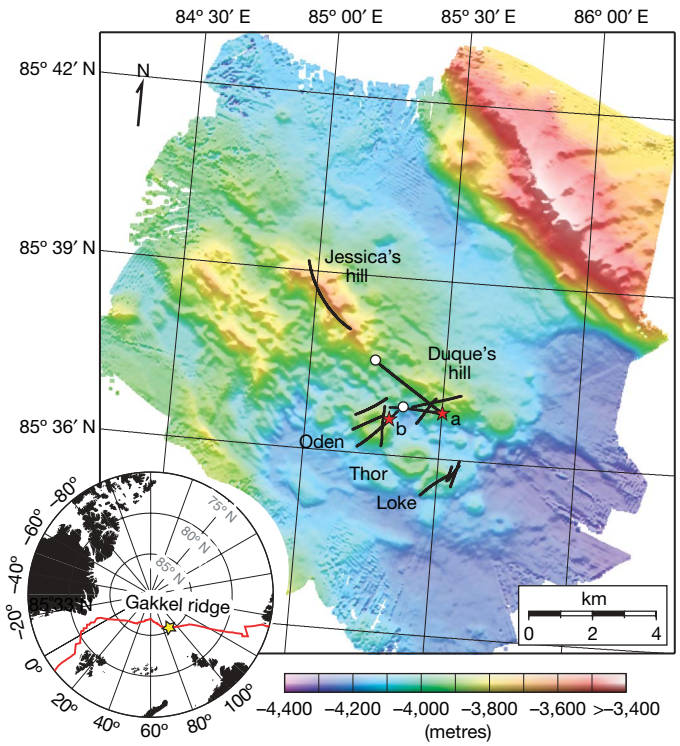


Figure 1 | Detailed bathymetry (30 m grid spacing) of the Gakkel ridge at 85° E in the Arctic Ocean. The inset map shows the location of the 85° E segment (yellow star) along the Gakkel ridge (red line) in the Arctic basin. The main panel shows illuminated, colour bathymetry of the 85° E segment acquired during the AGAVE expedition. The axial valley contains large numbers of distinctive, cratered volcanoes, including a cone on a fault terrace of the northern valley wall. Photographic bottom surveys were conducted along profiles shown as thin black lines on the map. Pyroclastic deposit samples were collected at sites shown by white circles, and the photographs shown in Fig. 2a, b were taken at the sites shown by the lettered red stars. Named features include two volcanic ridges in the centre of the axial valley (Jessica's hill and Duque's hill), and three cratered volcanoes along a ridge-parallel fissure to the south (Oden, Thor and Loke). The bathymetry data were plotted with Generic Mapping Tools[22].

[1]Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA. [2]Marine Science Institute, University of Texas at Austin, Port Aransas, Texas 78373, USA. [3]Swedish Polar Secretariat, 104 05 Stockholm, Sweden. [4]Alfred Wegener Institute for Polar and Marine Research, Bremerhaven 27570, Germany. [5]Johns Hopkins University, Baltimore, Maryland 21218, USA. [6]Göteborg University, Goteborg 40530, Sweden. [7]AIST, Tsukuba Central 7, 1-1-1 Higashi, Japan. [8]Ocean Research Institute, University of Tokyo, Minamidai, Nakano, Tokyo 164-8639, Japan. [9]Department of Geology and Geochemistry, Stockholm University, 106 91 Stockholm, Sweden.

crater and cluster on ridge-parallel faults or fissures. The type example is perhaps Oden volcano, which is ~300 m tall and ~1.5 km in diameter, and contains an ~50-m deep central crater ~500 m in diameter (Fig. 1).

A real-time fibre-optic connection allowed scientists aboard the icebreaker to 'fly' the CAMPER vehicle 2–5 m above the sea floor within the region of the suspected eruption and acquire photographic still imagery and high-definition video imagery of the volcanic terrain. The images reveal that the axial valley topography is blanketed with unconsolidated pyroclastic deposits up to 10 cm thick. The thickest deposits overlie the weathered and broken lava flows (Fig. 2a) on Jessica's hill and Duque's hill in the centre of the axial valley (Fig. 1), whereas the fresh (that is, unweathered, glassy surfaces with delicate ornamentations) lava flows on the Oden and Loke volcanoes are covered with a light dusting of material. Pyroclasts are piled atop pillows and other high-standing features, indicating deposition by fall rather than flow. Multiple falls are implied by spatial variations in the apparent age (colour) and thickness of the deposits. The maximum extent of the pyroclastic material is not known, because the deposits were observed to cover every portion of the sea floor that we surveyed (~20 linear kilometres within an ~5 × 10 km region).

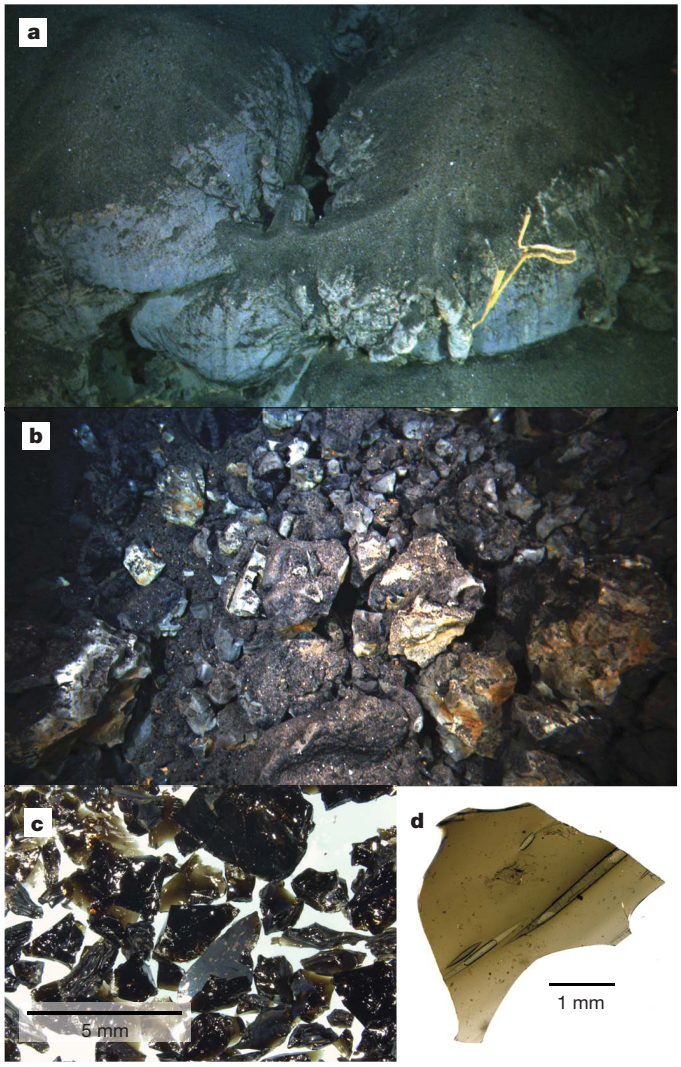


Figure 2 | Photographs of pyroclastic deposits. a, Frame grab from a high-definition video camera taken on the south side of Duque's hill (see Fig. 1 for location). About 10 cm (visually estimated and confirmed during sampling) of pyroclastic material is piled atop a high-standing, weathered, pillow feature. The exoskeleton of an as yet unidentified species of hexactinellid sponge[23] is visible in the foreground. **b**, High-definition video frame grab of talus blocks possibly representing ejecta from a vulcanian explosion on Oden volcano (see Fig. 1 for location). **c**, Glassy, granular, pyroclastic material. **d**, Bubble wall fragment from pyroclastic deposit.

A series of eight dives across the Oden and Loke volcanoes suggests that the ubiquitous cratered volcanoes may be source vents for pyroclastic eruptions, possibly including vulcanian explosions. These volcanoes contain most of the fresh lava flows observed in our survey, which consist primarily of pillows but also include ropey sheet flows, covering small areas (~100–200 m^2) on the top and around the outer edges of the constructional features. The mixture of young and old lava flows that we observed demonstrates that the high-acoustic-backscatter region imaged in 1999 does not represent a single, fresh lava flow[8]. The crater of Oden volcano is filled with weathered, basaltic talus but contains no visible fault scarps. The talus is covered by small amounts of pyroclasts, and the block sizes generally decrease on moving away from the crater centre, extending onto the outer slopes of the volcano (Fig. 2b). These observations are consistent with an interpretation of the talus blocks as country rock ejecta from a vulcanian explosion, which may also participate in crater formation, but at this point we cannot unequivocally exclude the possibility that the talus was formed by mass wasting.

About 0.8 kg of clasts was collected from two sites along our tracklines with a retractable slurp (suction) device mounted on the CAMPER vehicle. The samples consisted nearly entirely of juvenile clasts of glassy basalt (Fig. 2c) with a small (≪5%) component of crystalline and altered crystalline rock. The clasts are primarily angular fragments, many with cuspate surfaces, that range in size from 1 to 20 mm (the suction sampler does not preserve *in situ* sorting). The clasts contain minor olivine and plagioclase microphenocrysts, and have low (<5%) vesicularity. In addition, the deposits contain small amounts of limu o Pele[2,11], which are thin, glassy, bubble wall fragments, 3–20 mm across, with fluidal folded morphologies (Fig. 2d).

Large-volume pyroclastic deposits have been found in shallow water (500–1,750 m water depth) on the Azores Plateau[12–14], but the only previous evidence of pyroclastic material at water depths greater than 3,000 m (the critical depth for steam) is limited to small fragments recovered in sediment cores[15,16]. Hydrostatic pressure inhibits volume expansion, and below the steam threshold any explosive activity must result from magmatic volatiles rather than secondary surface effects. CO_2 is the most plausible exsolved volatile component for MOR basalts[15], and at 4,000 m water depth a CO_2 weight fraction of ~14% (ref. 3) is necessary to achieve the volume fraction of ~75% needed to fragment an erupting magma[17]. This value exceeds the maximum dissolved CO_2 concentrations measured in a MOR basalt (~1.4 wt% in a 'popping rock')[4] by an order of magnitude.

Volatiles that exsolve during magma ascent or decompression may coalesce to produce finite volumes of magma with gas volume fractions sufficient to trigger pyroclastic activity, even in magmas with primary volatile levels well below the fragmentation threshold. The nature of pyroclastic activity triggered by this process depends on the amount of volatiles and the depth at which fragmentation occurs. For example, if gas exsolution and expansion occur during the slow rise of an erupting dyke, and the rising bubbles coalesce in the upper few hundred metres of the crust (that is, slug flow), then Strombolian (bubble burst) activity may occur at the sea floor. The observation of bubble wall fragments in our pyroclastic samples is consistent with some level of Strombolian activity, but bubble coalescence and fragmentation in the shallow crust can only distribute clasts to maximum distances of ~20–40 m from the source vent[3], which is inconsistent with the widespread distribution of material over the >10-km^2 region observed in our survey.

A more energetic mechanism is required for depositing clasts more than a few tens of metres from the source vent, which is possible if fragmentation occurs deeper within the crust. The accumulation of a large volume of volatiles in the upper layer of a crustal magma chamber[18] provides the most plausible mechanism for deep fragmentation. Exsolved volatiles may accumulate in a chamber over long periods, and then discharge explosively when the roof is fractured during an eruption, spreading pyroclastic material over large

Table 1 | Variation in pyroclastic jet characteristics with magma chamber depth

Magma chamber roof depth (m)	Minimum CO_2 volume fraction in volatile-rich layer	Pyroclastic jet mixture density at vent (kg m^{-3})	Average jet exit velocity at vent (m s^{-1})	Plume rise height in water column (m)
1,000	0.6443	568	236	544
2,000	0.5647	462	343	956
3,000	0.5026	398	424	1,276
4,000	0.4528	355	490	1,532
5,000	0.4120	324	544	1,741
6,000	0.3779	302	591	1,916

Following the analysis presented in sections 2.4 and 2.5 of ref. 3, we calculate the minimum CO_2 volume fraction in a volatile-rich layer accumulating under the roof of a crustal magma chamber required for producing pyroclastic activity on the sea floor at a depth of 4,000 m. Magmas with the gas volume fractions shown in column 2 will fragment just before reaching the sea floor, producing very small deposits. However, if the CO_2 volume fraction in a volatile-rich layer is ~0.75, fragmentation occurs at the magma chamber depth, and a much more energetic eruption occurs as the gas accelerates during ascent to the sea floor, producing the approximate conditions shown in columns 3–5.

areas proportional to the chamber depth. For the range of parameters appropriate for our study area, we find that pyroclastic fountains may rise as high as 1–2 km in the water column if fragmentation occurs within a crustal chamber (Table 1).

These results provide a new perspective for interpreting the 1999 seismic swarm and volcanic event at the 85° E site. The seismic swarm began with extensional events, but after three months the earthquakes changed to sources with large volume changes (implosions)[6]. Large-volume-change events are rare at MORs, but they are consistent with the rapid evacuation of explosive material from a deep-lying magma chamber. The sequence of extensional earthquakes leading up to the implosions may have perturbed the stress field enough to fracture the chamber roof, thereby releasing pressurized magmatic volatiles. Rapid acceleration of decompressing volatiles may have triggered vulcanian explosions during the eruption[3], consistent with the talus distribution observed on Oden volcano. Multiple episodes of explosive volatile discharge over a prolonged period are required for producing the variations in apparent age and thickness of the deposits we observed, and we note that small-magnitude explosive acoustic signals were detected by local (ice-mounted) seismic networks at the eruption site more than two years after the 1999 seismic swarm[19]. Explosive volatile discharge has clearly been a widespread, and ongoing, process at the 85° E segment.

Our results raise new questions about volatile processes in ultraslow-spreading magmatic systems. More observations will be necessary to determine the ubiquity of pyroclastic activity at ultraslow spreading rates (<15–20 mm yr^{-1}, full rate), but from first principles there is reason to believe that ultraslow-spreading ridges may be especially conducive to the build-up and explosive discharge of volatile-rich magmatic foams. Long time intervals between eruptions should increase the quantity of volatiles that can be accumulated in a magma chamber, and if the global correlation between spreading rate and magma chamber depth extends to ultraslow rates, then volatile build-up will occur deep within the crust at high storage pressures. Our results add to the growing body of evidence that ultraslow-spreading ridges host unique modes of crustal accretion and tectonic extension[20,21], and motivate continuing efforts to solve the technical and logistical issues that have impeded scientific access to these unique geological environments.

Received 19 December 2007; accepted 6 May 2008.

1. White, J. D. L., Smellie, J. L. & Clague, D. A. in *Explosive Subaqueous Volcanism* (eds White, J. D. L., Smellie, J. L. & Clague, D. A.) 1–24 (American Geophysical Union, Washington DC, 2003).
2. Hon, K., Heliker, C. C. & Kjargaard, J. I. Limu o Pele: A new kind of hydroclastic tephra from Kilauea Volcano. *Hawai'i Geol. Soc. Am. Abst. Prog.* **20,** A112–A113 (1988).
3. Head, J. W. I. & Wilson, L. Deep submarine pyroclastic eruptions: theory and predicted landforms and deposits. *J. Volcanol. Geotherm. Res.* **121,** 155–193 (2003).
4. Javoy, M. & Pineau, F. The volatiles record of a 'popping' rock from the Mid-Atlantic Ridge at 14°N: Chemical and isotopic composition of gas trapped in the vesicles. *Earth Planet. Sci. Lett.* **107,** 598–611 (1991).
5. Sella, G. F., Dixon, T. H. & Mao, A. REVEL: A model for recent plate velocities from space geodesy. *J. Geophys. Res.* **107,** doi:10.1029/2000JB000033 (2002).
6. Mueller, C. & Jokat, W. Seismic evidence for volcanic activity discovered in central Arctic. *Eos* **81,** 265–269 (2000).
7. Tolstoy, M., Bohnenstiehl, D. R., Edwards, M. H. & Kurras, G. J. Seismic character of volcanic activity at the ultraslow-spreading Gakkel Ridge. *Geology* **29,** 1139–1142 (2001).
8. Edwards, M. H. *et al.* Evidence of recent volcanic activity on the ultraslow-spreading Gakkel ridge. *Nature* **409,** 808–812 (2001).
9. Edmonds, H. N. *et al.* Discovery of abundant hydrothermal venting on the ultraslow-spreading Gakkel ridge in the Arctic Ocean. *Nature* **421,** 252–256 (2003).
10. Graham, D. W. *et al.* Helium-3, methane, and manganese in water column hydrothermal plumes along the ultra-slow spreading Gakkel Ridge, Arctic Ocean. *Ofioliti* **31,** 234–235 (2006).
11. Clague, D. A., Davis, A. S., Bischoff, J. L., Dixon, H. E. & Geyer, R. Lava bubble-wall fragments formed by submarine hydrovolcanic explosions on Loihi Seamount and Kilauea Volcano. *Volcano Bull. Volcanol.* **61,** 437–449 (2000).
12. Eissen, J.-P., Fouquet, Y., Hardy, D. & Ondreas, H. in *Explosive Subaqueous Volcanism* (eds White, J. D. L., Smellie, J. L. & Clague, D. A.) 143–166 (American Geophysical Union, Washington DC, 2003).
13. Fouquet, Y. *et al.* Extensive volcaniclastic deposits at the Mid-Atlantic Ridge axis; results of deep-water basaltic explosive volcanic activity? *Terra Nova* **10,** 280–286 (1998).
14. Hekinian, R. *et al.* Deep sea explosive activity on the Mid-Atlantic Ridge near 34° 50′ N: Magma composition, vesicularity, and volatile content. *J. Volcanol. Geotherm. Res.* **98,** 49–77 (2000).
15. Clague, D. A., Davis, A. S. & Dixon, J. E. in *Explosive Subaqueous Volcanism* (eds White, J. D. L., Smellie, J. L. & Clague, D. A.) 111–128 (American Geophysical Union, Washington DC, 2003).
16. Clague, D. A., Uto, K., Satake, K. & Davis, A. S. in *Hawaiian Volcanoes: Deep Underwater Perspectives* (eds Takahashi, E., Lipman, P. W., Garcia, M. O., Naka, J. & Aramaki, S.) 65–84 (American Geophysical Union, Washington DC, 2002).
17. Sparks, R. S. J. The dynamics of bubble formation and growth in magmas: A review and analysis. *J. Volcanol. Geotherm. Res.* **3,** 1–37 (1978).
18. Bottinga, Y. & Javoy, M. Mid-ocean ridge basalt degassing: Bubble nucleation. *J. Geophys. Res.* **95,** 5125–5131 (1990).
19. Schlindwein, V., Müller, C. & Jokat, W. Seismoacoustic evidence for volcanic activity on the ultraslow-spreading Gakkel Ridge, Arctic Ocean. *Geophys. Res. Lett.* **32,** doi:10.1029/2005GL023767 (2005).
20. Dick, H. J. B., Lin, J. & Schouten, H. An ultraslow-spreading class of ocean ridge. *Nature* **426,** 405–412 (2003).
21. Michael, P. J. *et al.* Magmatism and amagmatic seafloor generation at the ultraslow-spreading Gakkel ridge, Arctic Ocean. *Nature* **423,** 956–961 (2003).
22. Wessel, P. & Smith, W. H. F. Free software helps map and display data. *Eos* **72,** 441 (1991).
23. Shank, T. M. *et al.* Biological and geological characteristics of the Gakkel Ridge. *Eos* 88, abstr. OS41C–08 (2007).

Acknowledgements We thank D. Clague and J. Head for reviews that improved the final manuscript, the Advanced Imaging Laboratory at WHOI for technical support and the crew of icebreaker *Oden* for technical support at sea. This research was funded by the National Aeronautics and Space Administration, the National Science Foundation, and the Woods Hole Oceanographic Institution.

Author Contributions R.A.S. was the chief scientist of the Arctic Gakkel Vents Expedition and wrote the paper. H.S., T.M.S, S.H. and C.W. collected the dive imagery. T.M.S., S.H., C.W. and A.S. analysed the dive imagery. M.J. processed the bathymetric data. H.N.E., C.K., U.H., E.H., M.J., B.L., J.L., C.M., K.N., T.S., V.S, C.S., M.T., L.U. and P.W. provided technical and scientific support at sea. All authors discussed the results and provided input to the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.A.S. (rsohn@whoi.edu).

LETTERS

Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution

Brian T. Wilhelm[1]*†, Samuel Marguerat[1]*†, Stephen Watt[1]†, Falk Schubert[1]†, Valerie Wood[1], Ian Goodhead[1]†, Christopher J. Penkett[1]†, Jane Rogers[1] & Jürg Bähler[1]†

Recent data from several organisms indicate that the transcribed portions of genomes are larger and more complex than expected, and that many functional properties of transcripts are based not on coding sequences but on regulatory sequences in untranslated regions or non-coding RNAs[1–9]. Alternative start and polyadenylation sites and regulation of intron splicing add additional dimensions to the rich transcriptional output[10,11]. This transcriptional complexity has been sampled mainly using hybridization-based methods under one or few experimental conditions. Here we applied direct high-throughput sequencing of complementary DNAs (RNA-Seq), supplemented with data from high-density tiling arrays, to globally sample transcripts of the fission yeast *Schizosaccharomyces pombe*, independently from available gene annotations. We interrogated transcriptomes under multiple conditions, including rapid proliferation, meiotic differentiation and environmental stress, as well as in RNA processing mutants to reveal the dynamic plasticity of the transcriptional landscape as a function of environmental, developmental and genetic factors. High-throughput sequencing proved to be a powerful and quantitative method to sample transcriptomes deeply at maximal resolution. In contrast to hybridization, sequencing showed little, if any, background noise and was sensitive enough to detect widespread transcription in >90% of the genome, including traces of RNAs that were not robustly transcribed or rapidly degraded. The combined sequencing and strand-specific array data provide rich condition-specific information on novel, mostly non-coding transcripts, untranslated regions and gene structures, thus improving the existing genome annotation. Sequence reads spanning exon–exon or exon–intron junctions give unique insight into a surprising variability in splicing efficiency across introns, genes and conditions. Splicing efficiency was largely coordinated with transcript levels, and increased transcription led to increased splicing in test genes. Hundreds of introns showed such regulated splicing during cellular proliferation or differentiation.

To analyse the *S. pombe* transcriptome at the best possible resolution, we used Illumina 1G to sequence directly cDNA synthesized from poly(A)-enriched RNA. This approach kept the proportion of sequence reads from ribosomal RNA low (<10%) without biasing against messenger RNAs with short poly(A) tails[12]. We obtained >23 million reads of an average length of 39.1 base pairs (bp), representing ~60 genome lengths, from cells proliferating exponentially in rich medium. In addition, we acquired >99 million reads of transcriptomes from five stages of meiotic differentiation, representing an additional ~190 genomes (Supplementary Table 1). Sequence reads were mapped back to both the spliced and the unspliced reference genome[13] to determine the numbers of reads hitting each genomic base-pair position. Approximately 60% of all reads specifically mapped to one genomic region over 100% of their sequence, whereas >85% of the reads uniquely mapped over 90% of their

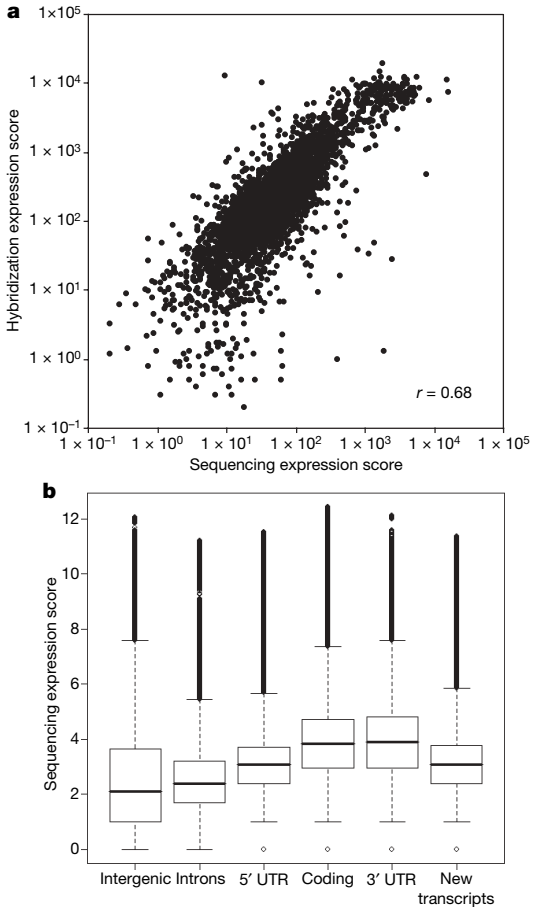


Figure 1 | Quantitation of sequence coverage. a, Scatterplot comparing gene-expression scores based on Affymetrix expression-chip hybridization signals (*y* axis) with gene-expression scores based on high-throughput sequencing (*x* axis). The dynamic range of hybridization signals is limited by the scanner. The corresponding Pearson correlation is shown at the bottom right. **b**, Box-and-whisker plots (in which the whiskers denote the 5th and 95th quantiles) of $\log_2$-transformed numbers of sequence reads per nucleotide for the following genomic regions: all intergenic sequences, introns, coding sequences, 5′ and 3′ UTRs (based on sequencing), and newly identified transcripts (based on sequencing and tiling chips). Diamonds represent data outside of the quantiles.

[1]Cancer Research UK Fission Yeast Functional Genomics Group, Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1HH, UK. †Present addresses: Institut de Recherche en Immunologie et en Cancérologie (IRIC), Montreal, H3C 3J7, Canada (B.T.W.); Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK (S.M., S.W., F.S. and J.B); School of Biological Sciences, University of Liverpool, L69 7ZB, UK (I.G.); EMBL-European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK (C.J.P.). *These authors contributed equally to this work.

sequence. The remaining reads either mapped to repeated sequences or were of poor quality. RNA expression levels determined from sequence-read numbers strongly correlated with those determined from hybridization signals, indicating that sequencing provides quantitative data on transcript levels (Fig. 1a).

The 5% of transcripts present at the lowest steady-state levels in rapidly proliferating cells[12] accumulated ~777 sequence-read hits and 94.9% coverage on average, indicating that the transcriptome was sampled deeply enough to detect even genes with low expression levels. We modelled sequencing depth for rapidly proliferating cells: given the expression scores for all annotated genes, the model predicts that 99% of these genes have >50% sequence-read coverage (Supplementary Fig. 1). In agreement with this prediction, we obtained >50% sequence-read coverage for 99.3% of all annotated genes. The 41 genes with <50% coverage included 20 transposon-related long terminal repeats and 13 dubious genes or pseudogenes (Supplementary Table 2). Using cDNA microarrays, only 80–90% of genes yield measurable signals in proliferating cells[14], whereas the remaining genes are only highly expressed under specific conditions such as meiosis or stress[15,16]. These data suggest that the sequencing approach is sensitive enough to detect basal 'transcriptional noise' from genes that are not actively expressed.

As expected, intergenic regions were hit by fewer sequence reads than coding regions (Figs 1b and 2a). However, we obtained sequence data from ~94% and >99% of the nuclear and mitochondrial genomes, respectively, suggesting that almost the entire genome is transcribed to some degree, consistent with the considerable overlap and complexity among different transcripts reported for other eukaryotes[9]. Reverse transcription followed by polymerase chain reaction (RT–PCR) controls verified that even intergenic regions with poor sequence-read coverage reflect expressed RNAs rather than technical noise from spurious sequences (Supplementary Fig. 2). Thus, our sequence data provide direct evidence for widespread transcription; it has been suggested that as much as 90% of all RNA polymerase II (Pol II) initiation events represent transcriptional noise[17]. Taken together, unlike for hybridization-based approaches, sequencing appears to produce little or no background noise, and the dynamic range of detected transcripts is only limited by sequencing depth.

To verify and compare the sequence data with an established platform, we used Affymetrix chips containing 25-mer probes tiled at ~20-nucleotide intervals across both strands of the *S. pombe* genome. We interrogated transcriptomes under a wide range of conditions (Supplementary Table 1), thus independently sampling gene expression at lower resolution but with strand-specific information (Fig. 2a).

The combined sequence and hybridization data revealed hundreds of novel transcribed regions. To distinguish between separate transcripts and extensions to known gene structures, we analysed tiling-chip data from a *prp2* splicing-factor mutant[18] along with sequence 'trans-reads' spanning unannotated splice junctions (Figs 2a and 3d). Combined with manual curation, these analyses helped to refine annotated gene structures, including 75 revisions of protein-coding

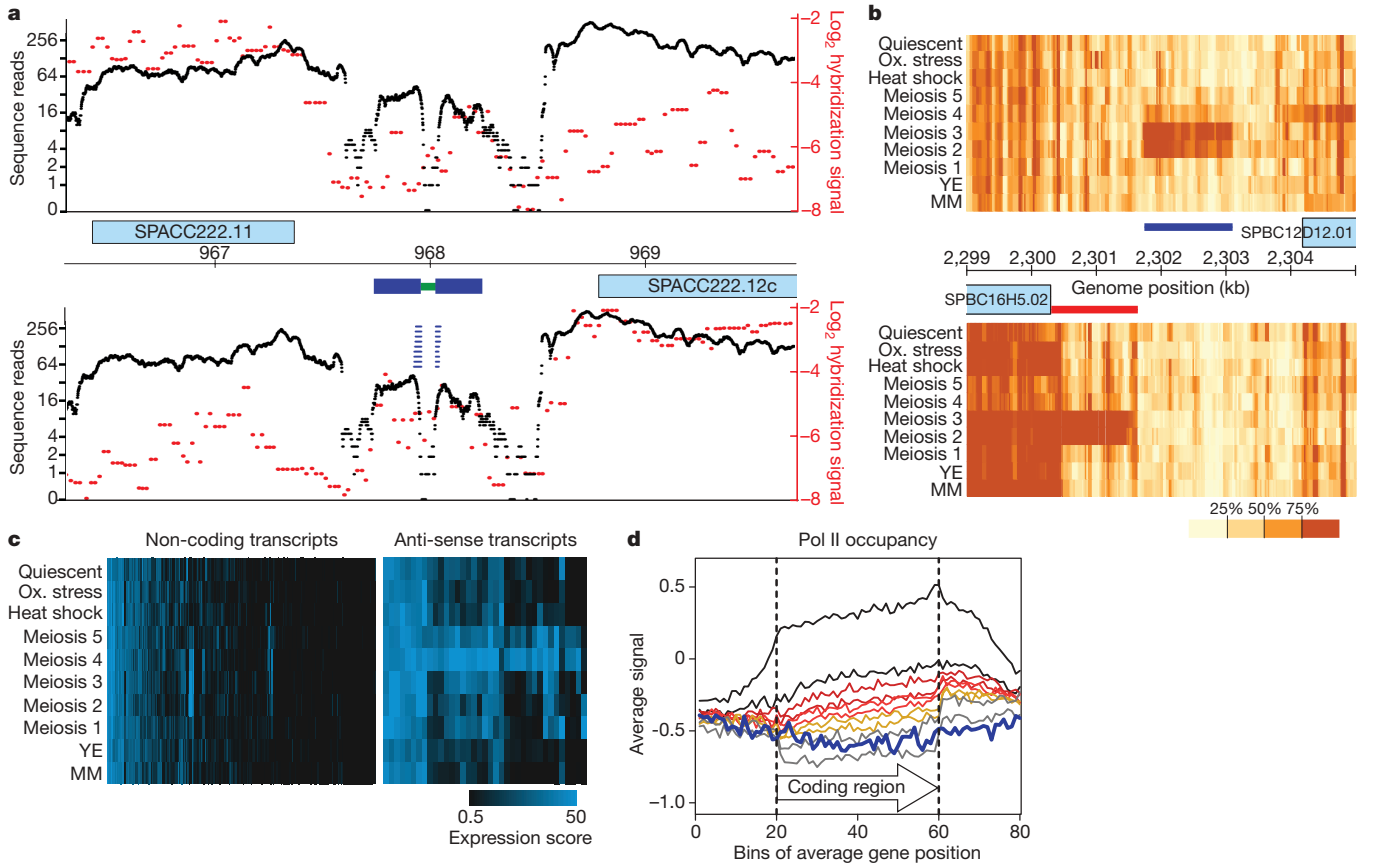


Figure 2 | Analyses of novel transcripts. **a**, Plot depicting numbers of sequence reads (log-scale, black) and tiling-chip hybridization signals (log_2, red) across the genomic region indicated in the centre (coordinates in kb) for forward (top) and reverse (bottom) strands. The sequence data are not strand-specific. A novel non-coding transcript (dark blue) containing an intron (green) is indicated in the middle. The nine trans-reads across the exon–exon junction are indicated as broken blue bars. **b**, Tiling-chip hybridization signals (in which the strength of colour reflects the signal-distribution quartile) across the genomic region shown in the centre for forward (top) and reverse (bottom) strands, with rows reflecting ten experimental conditions (Supplementary Table 1). Rapid proliferation was sampled in rich (YE) and minimal (MM) media. Blue bar, novel meiosis-specific transcript; red bar, alternate meiosis-specific 5′ UTR. **c**, Hierarchical clustering of non-coding and anti-sense transcripts by their tiling-chip expression scores across multiple conditions as in **b**. **d**, Average Pol II occupancy across coding regions for genes with lowest to highest mRNA levels (grey to black via red to yellow shades). The average profile of the novel transcripts is shown as a thick blue line.

regions and identification of ~20 new introns in known genes. Conservative data analysis also revealed 453 novel transcripts, only 26 of which seemed to be coding for small proteins (<150 amino acids); 37 of the apparently non-coding transcripts overlapped known genes in the anti-sense direction (Supplementary Table 3). The 427 non-coding RNAs showed an average length of ~825 nucleotides and a GC content that was similar to the 135 annotated non-coding RNAs but higher than for intergenic regions overall (33.0% versus 30.6%; $P < 2 \times 10^{-16}$, Wilcoxon test). The non-coding RNAs included the elusive, recently discovered Ter1 telomerase RNA[19,20], which was induced during meiosis (SPNCRNA.214; Supplementary Table 3). Expression of 14 non-coding RNAs was independently confirmed by RT–PCR (Supplementary Fig. 3). This analysis revealed bi-directional transcription across all tested regions, including the well-characterized *nmt1* gene, although most regions showed more transcripts from one strand. Given the ubiquitous transcription throughout the genome, the novel transcripts described here probably only hint at the true level of transcriptional complexity.

Sequence-read numbers across the newly identified transcribed regions were lower than numbers across annotated coding regions (Fig. 1b). Only 13 of the novel transcripts were evident from the tiling-chip data in proliferating cells, whereas another 79 were only substantially expressed under specific conditions, most notably during meiosis or quiescence (Fig. 2b, c and Supplementary Table 3). The antisense RNAs were particularly enriched for highly regulated transcripts, many of which peaked during the meiotic divisions (Fig. 2c).

To test whether some of the newly identified regions reflect cryptic transcripts that are degraded in the nucleus, we analysed RNA isolated from an *rrp6* mutant defective in nuclear exosome function[21,22]: 36 of the novel transcripts were more highly expressed in this mutant such that they became evident also on tiling chips (Supplementary Table 3). These data raised the possibility that many newly identified regions are strongly transcribed but rapidly degraded by different surveillance systems[21]. To test this hypothesis, we globally measured Pol II occupancy (reflecting transcriptional activity[12]). Overall, Pol II occupancy across the new regions was comparable to the location of 10–20% of genes with the lowest levels of transcription (Fig. 2d). We conclude that most newly identified regions were not robustly expressed in proliferating cells, but that the sequencing approach was sufficiently sensitive to detect transcriptional traces below the detection limit of hybridization-based approaches.

The combined sequence and hybridization data provided a rich source to analyse transcript structures at maximal resolution. High densities of overlapping transcripts can confound the sequence data, and decreasing read-numbers towards the 5′ ends, reflecting oligo(dT) priming (Figs 1b and 3a), render it difficult to determine accurately transcript lengths of long genes. The hybridization data are less affected by these issues because they distinguish transcriptional direction and do not show any 5′ bias (Fig. 3a and Supplementary Fig. 4). Together, the two approaches provided complementary data on untranslated regions (UTRs) for most *S. pombe* genes (Supplementary Table 4). For many other genes, which were mostly expressed at low levels and did not pass our confidence cutoffs, the

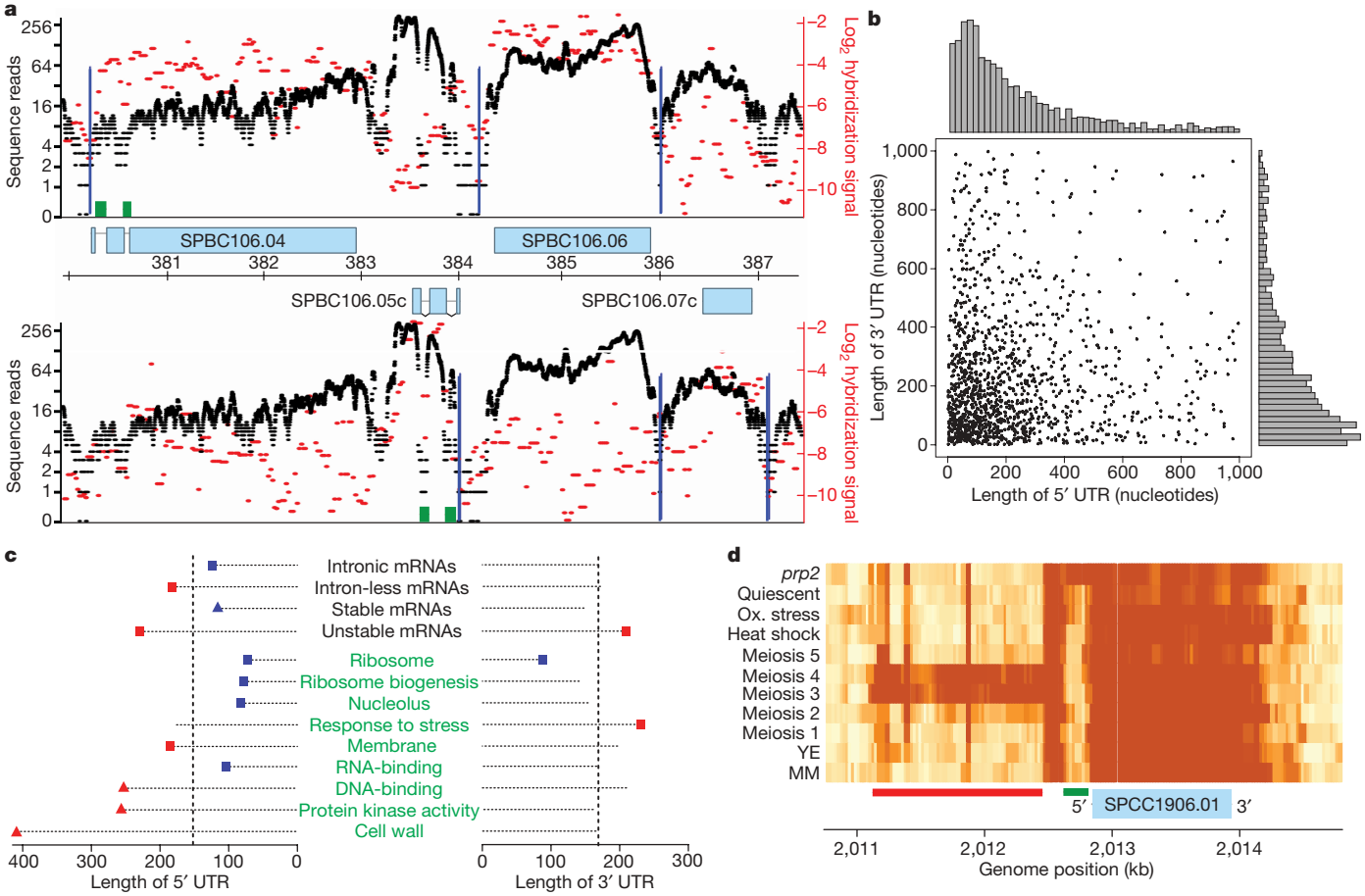


Figure 3 | Analyses of transcript structures. **a**, Plot depicting sequence reads and tiling-chip signals as in Fig. 2a. Vertical dark blue lines, transcription start and end sites determined by sequencing; green boxes, introns. **b**, Scatterplot and histograms showing length distributions of 3′ and 5′ UTRs based on tiling-chip data. **c**, Transcripts with significantly larger (red) or smaller (blue) UTRs for selected mRNA properties (top) or GO categories (bottom, green), based on tiling-chip data (triangles) or on tiling-chip and sequence data (squares) (Wilcoxon test, $P < 0.05$, Hochberg-adjusted for multiple tests). Vertical dashed lines: median UTR lengths. **d**, Tiling-chip hybridization signals as in Fig. 2b, showing a novel intron that is not spliced in the *prp2* mutant (green bar) and an alternate 5′ UTR (red bar).

UTRs could be mapped by visual inspection. UTRs determined by hybridization or sequencing showed good agreement with each other and also with the previously known UTRs (Supplementary Table 5). The median 5′- and 3′-UTR lengths determined by hybridization were 152 and 169 nucleotides, respectively, with a mean combined length of 465 nucleotides (Fig. 3b). Thus, the UTRs of fission yeast are substantially larger than those of budding yeast, which show a mean combined length of 211 nucleotides[5].

We compared UTR-length distributions for different functional categories (Fig. 3c). The most stable transcripts[12] had short 5′ UTRs, whereas the least stable transcripts had long 5′ and 3′ UTRs, which may contain regulatory signals for RNA turnover. An analysis of Gene Ontology (GO) categories with significantly longer or shorter UTRs (Fig. 3c) uncovered similarities to budding yeast[5]. For example, transcripts encoding protein kinases and membrane proteins had long 5′ UTRs, whereas ribosome-biogenesis genes had short 5′ UTRs in both yeasts, indicating that UTR-length distributions show some conservation in these distantly related yeasts.

Sampling UTR lengths under different conditions allowed detection of transcript-size regulation (Supplementary Table 4). Our data confirmed the known transcripts with alternate start sites or polyadenylation sites produced from *cig2* and *wos2*, respectively[23,24]. Using a conservative approach, we identified 27 additional transcripts with alternate start sites during meiosis or stress (Figs 2b and 3d, and Supplementary Table 6). Alternate polyadenylation sites were more abundant, affecting ~187 transcripts (Supplementary Table 6). Transcription-termination sites were generally less well defined than start sites and also varied across different conditions (Fig. 3d and Supplementary Fig. 5).

The resolution of the tiling chips was limiting to analyse splicing owing to the small size of most introns (<100 nucleotides). The sequence data, however, provided unprecendented insights into splicing of the 45.4% intronic genes of *S. pombe*[13]. Both unspliced and spliced transcripts were present in the total RNA preparations; accordingly, we also obtained reads covering introns, albeit at lower numbers than for exons (Figs 1b and 3a). Importantly, sequencing provided direct evidence for splicing owing to 'trans-reads' spanning exon–exon junctions, thus confirming ~93% of predicted introns and hugely reducing unsupported gene structures. We found no evidence for the existence of alternate splicing in *S. pombe*.

To estimate splicing efficiencies, we determined normalized numbers of sequence reads spanning exon–exon and corresponding exon–intron junctions for all introns (Supplementary Table 7). This calculation of splicing efficiency exploits relative read numbers and is therefore internally normalized for expression levels and sequencing depth. Median numbers of spliced transcripts were only ~2-fold higher than numbers of corresponding unspliced transcripts, suggesting a surprisingly large cellular portion of unprocessed mRNAs (Supplementary Table 7). Average splicing efficiency was similar for different intron positions within genes (Supplementary Fig. 6). Splicing efficiency strongly varied, however, among different genes and conditions. A conservative analysis uncovered 254 genes (314 introns) that were more efficiently spliced during meiotic differentiation than in proliferating cells (Supplementary Table 8). These genes included 9 of 12 known meiotically spliced genes[25], whereas the 3 remaining genes showed increased meiotic splicing below our cutoff. Such 'regulated' splicing was evident in all five differentiation stages tested, but was most prevalent during meiotic prophase and nuclear divisions (Fig. 4a). In some genes all introns showed regulated splicing, whereas in others only selected introns were regulated (Supplementary Table 8)—a finding that was robust to lowering the cutoff. The median proportion of introns per gene showing regulated splicing was 50%, and regulated splicing showed no preference for specific intron positions.

The surprisingly large, yet conservative, list of genes with increased meiotic splicing was highly enriched for genes showing increased transcript levels during meiosis[26] ($P \sim 2 \times 10^{-20}$, hypergeometric test). Coordinated increases of meiotic gene expression and splicing

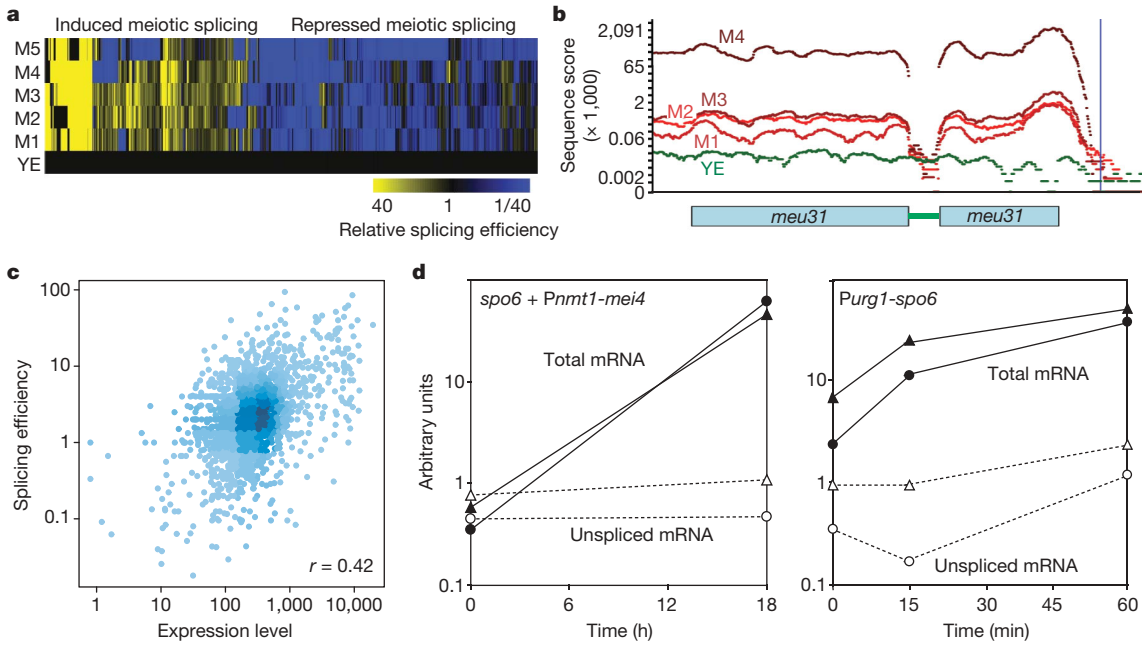


Figure 4 | Dynamics of splicing efficiency reflects transcription. **a**, Hierarchical clustering of introns by their splicing efficiency in five stages of meiotic differentiation (M1 to M5) relative to their splicing efficiency during rapid proliferation (YE). **b**, Plot depicting log-scale numbers of sequence reads normalized for sequencing depth across *meu31* (intron depicted as green bar), colour-coded by experimental condition. Numbers of trans-reads across the exon–exon junction range from 0 (YE) to 490 (M3). **c**, Scatterplot comparing median splicing efficiency for intron-containing genes with mRNA levels based on expression-chip hybridization signals. Shades of blue reflect the gene density, and Pearson correlation is shown at the bottom right. **d**, RT–PCR data to quantify splicing of *spo6* transcript as a function of transcription. Left: RNA levels before and 18 h after overexpression of Mei4 using the *nmt1* promoter (P*nmt1*)[27]; right: RNA levels before and up to 1 h after direct overexpression of *spo6* using the *urg1* promoter (P*urg1*)[30]. Data from primers within exons (solid) or across exon–intron junctions (dashed) are shown for two different exons or junctions, respectively.

were also directly evident from the sequence data (Fig. 4b). Moreover, meiotic transcripts showed similar profiles for gene expression and splicing efficiency during meiosis (Supplementary Fig. 7). A reciprocal analysis uncovered 478 genes (559 introns) that were more efficiently spliced in proliferating cells than during meiosis (Fig. 4a and Supplementary Table 8). This list was enriched for genes highly expressed in proliferating cells[16], including ribosomal-protein genes ($P<2\times10^{-7}$, hypergeometric test). These data suggest that increased transcription can promote splicing. Indeed, splicing efficiency was significantly correlated with mRNA levels (Fig. 4c). Moreover, a functional analysis revealed widespread relationships between expression levels and splicing efficiency in proliferating cells (Supplementary Table 9). For example, highly expressed genes, such as those repressed during stress[15], or conserved genes[16] were more efficiently spliced than genes induced during stress or than *S. pombe*-specific genes.

To test directly whether increased transcription can lead to increased splicing, we activated transcription of the meiotically spliced *spo6* and *spn7* genes, either by placing them under the control of an ectopic regulatable promoter or by overexpressing the transcription factor Mei4, which activates *spo6* and *spn7* (ref. 27) and has been implicated in the regulation of meiotic splicing[28]. The proportion of spliced transcripts increased after activating transcription, using either the ectopic or the native transcription factor (Fig. 4d; Supplementary Fig. 7). We conclude that activation of transcription itself is sufficient to promote splicing during meiosis, without the specific need for the meiotic factor Mei4. This finding raises the possibility that transcriptional and splicing efficiencies are mechanistically linked. Taken together, our results reveal a surprising genome-wide regulation of splicing, largely reflecting transcript levels during proliferation or differentiation. These data point to a global and condition-specific coupling between splicing efficiency and transcription, which may help to optimize and streamline gene expression programmes.

METHODS SUMMARY

Strains and experimental conditions are listed in Supplementary Table 1. cDNA for sequencing and array hybridization was prepared using oligo(dT) or random primers, respectively. For sequencing, fragment sizes of 120–170 bp were attached to the FlowCell at an average concentration of 3 pM, amplified isothermally, and sequenced using Solexa reversible-terminator chemistry on the Illumina Genome Analyser. Sequence reads were mapped to the reference genome using BLAT. Analyses of tiling-chip data were based on the Bioconductor package 'tilingArray'[29].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 17 March; accepted 15 April 2008.
Published online 18 May 2008.

1. Yamada, K. *et al.* Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302,** 842–846 (2003).
2. Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306,** 2242–2246 (2004).
3. Stolc, V. *et al.* A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* **306,** 655–660 (2004).
4. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309,** 1559–1563 (2005).
5. David, L. *et al.* A high-resolution map of transcription in the yeast genome. *Proc. Natl Acad. Sci. USA* **103,** 5320–5325 (2006).
6. Li, L. *et al.* Genome-wide transcription analyses in rice using tiling microarrays. *Nature Genet.* **38,** 124–129 (2006).
7. The Encode Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
8. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316,** 1484–1488 (2007).
9. Kapranov, P., Willingham, A. T. & Gingeras, T. R. Genome-wide transcription and the implications for genomic organization. *Nature Rev. Genet.* **8,** 413–423 (2007).
10. Blencowe, B. J. Alternative splicing: new insights from global analyses. *Cell* **126,** 37–47 (2006).
11. Hughes, T. A. Regulation of gene expression by alternative untranslated regions. *Trends Genet.* **22,** 119–122 (2006).
12. Lackner, D. H. *et al.* A network of multiple regulatory layers shapes gene expression in fission yeast. *Mol. Cell* **26,** 145–155 (2007).
13. Wood, V. *et al.* The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415,** 871–880 (2002).
14. Lyne, R. *et al.* Whole-genome microarrays of fission yeast: characteristics, accuracy, reproducibility, and processing of array data. *BMC Genomics* **4,** 27 (2003).
15. Chen, D. *et al.* Global transcriptional responses of fission yeast to environmental stress. *Mol. Biol. Cell* **14,** 214–229 (2003).
16. Mata, J. & Bähler, J. Correlations between gene expression and gene conservation in fission yeast. *Genome Res.* **13,** 2686–2690 (2003).
17. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Struct. Mol. Biol.* **14,** 103–105 (2007).
18. Potashkin, J., Li, R. & Frendewey, D. Pre-mRNA splicing mutants of *Schizosaccharomyces pombe*. *EMBO J.* **8,** 551–559 (1989).
19. Leonardi, J., Box, J. A., Bunch, J. T. & Baumann, P. TER1, the RNA subunit of fission yeast telomerase. *Nature Struct. Mol. Biol.* **15,** 26–33 (2008).
20. Webb, C. J. & Zakian, V. A. Identification and characterization of the *Schizosaccharomyces pombe* TER1 telomerase RNA. *Nat. Struct. Mol. Biol.* **15,** 34–42 (2008).
21. Bickel, K. S. & Morris, D. R. Silencing the transcriptome's dark matter: mechanisms for suppressing translation of intergenic transcripts. *Mol. Cell* **22,** 309–316 (2006).
22. Harigaya, Y. *et al.* Selective elimination of messenger RNA prevents an incidence of untimely meiosis. *Nature* **442,** 45–50 (2006).
23. Borgne, A., Murakami, H., Ayté, J. & Nurse, P. The G1/S cyclin Cig2p during meiosis in fission yeast. *Mol. Biol. Cell* **13,** 2080–2090 (2002).
24. Munoz, M. J., Daga, R. R., Garzon, A., Thode, G. & Jimenez, J. Poly(A) site choice during mRNA 3′-end formation in the *Schizosaccharomyces pombe wos2* gene. *Mol. Genet. Genomics* **267,** 792–796 (2002).
25. Averbeck, N., Sunder, S., Sample, N., Wise, J. A. & Leatherwood, J. Negative control contributes to an extensive program of meiotic splicing in fission yeast. *Mol. Cell* **18,** 491–498 (2005).
26. Mata, J., Lyne, R., Burns, G. & Bähler, J. The transcriptional program of meiosis and sporulation in fission yeast. *Nature Genet.* **32,** 143–147 (2002).
27. Mata, J., Wilbrey, A. & Bähler, J. Transcriptional regulatory network for sexual differentiation in fission yeast. *Genome Biol.* **8,** R217 (2007).
28. Malapeira, J. *et al.* A meiosis-specific cyclin regulated by splicing is required for proper progression through meiosis. *Mol. Cell. Biol.* **25,** 6330–6337 (2005).
29. Huber, W., Toedling, J. & Steinmetz, L. M. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics* **22,** 1963–1970 (2006).
30. Watt, S. *et al. urg1*: a uracil-regulatable promoter system for fission yeast with short induction and repression times. *PLoS ONE* **3,** e1428 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. Gould and M. Yamamoto for strains, W. Huber and R. Durbin for advice, and J. Mata, W. Huber, V. Pancaldi, D. Stemple, J.-R. Landry and D. Lackner for comments on the manuscript. B.T.W. was supported by Sanger Postdoctoral and Canadian NSERC fellowships, and S.M. by a fellowship for Advanced Researchers from the Swiss National Science Foundation. This research was funded by Cancer Research UK grant number C9546/A6517, by the Wellcome Trust, and by DIAMONDS, an EC FP6 Lifescihealth STREP (LSHB-CT-2004-512143).

Author Contributions B.T.W., S.M. and J.B. designed and supervised the research and discussed the results; S.W. performed most experiments with help of B.T.W. and S.M.; B.T.W. and S.M. analysed the data with help of F.S., V.W., C.J.P. and J.B.; I.G. and J.R. helped with sequencing; and J.B. drafted the manuscript.

Author Information Raw data are available from ArrayExpress (http://www.ebi.ac.uk/arrayexpress) under accession numbers E-MTAB-5 (sequence data) and E-MTAB-18 (array data). Transcript data-plots are available from our TranscriptomeViewer at http://www.sanger.ac.uk/PostGenomics/S_pombe/. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.B. (jurg@sanger.ac.uk).

METHODS

cDNA preparation for high-throughput sequencing. All cDNA samples for Illumina were prepared by first treating ~1 mg of total RNA for 30 min with amplification-grade RNase-free DNase (Invitrogen), according to the manufacturer's protocols. Poly(A)-enriched RNA was then prepared using an oligo(dT) selection kit (Oligotex Direct mRNA miniKit, Qiagen). The resulting RNA was converted to double-stranded cDNA using a cDNA synthesis kit (Superscript choice system for cDNA synthesis, Invitrogen), primed by an oligo(dT) primer. RNA samples from the pooled meiotic time points were subjected to amplification by *in vitro* transcription (IVT) after a poly(A)-enrichment step as described above.

DNA libraries were prepared following the manufacturer's instructions (Illumina). DNA was sheared by nebulization, followed by simultaneous end-repair and phosphorylation using T4 DNA polymerase, Klenow fragment of DNA polymerase I and T4 PNK. DNA recovery was performed after each stage using QIAquick PCR purification columns (Qiagen). These repaired fragments were 3′-adenylated using Klenow exonuclease-minus (Illumina) and were purified using a MinElute PCR purification column (Qiagen). Illumina adaptors were ligated to the adenylated ends of the fragments and gel-purified on a 2% TAE (Tris-acetate-EDTA)-agarose gel (Certified Low-Range Ultra Agarose, Biorad), stained using ethidium bromide and visualized on a Dark Reader (Clare Chemical). A range of fragment sizes (120–170 bp) was excised from the gel and extracted using a QIAquick gel extraction kit. Seventeen rounds of PCR amplification were performed using primers complementary to the previously ligated adaptors and compatible to oligonucleotides attached to the FlowCell. DNA was recovered using a QIAquick PCR purification column. DNA was subsequently diluted to a working concentration of 10 nM in TE (Tris-EDTA) after quantification on a Nanodrop-1000 spectrophotometer.

Sequencing data processing and expression scores. FASTQ files of sequence reads were converted into FASTA files, and were filtered to remove sequences <15 bp after trimming the sequence from the position of the first *N*. All remaining FASTA sequences were matched back to the *S. pombe* genome using BLAT (tilesize 8, oneoff 1) in parallel on the Sanger Institute computer farm. All FASTA reads were also matched back as above to a spliced genome with all known or predicted intron sequences removed. The result files of matches to the spliced and unspliced genomes were compiled into a complete and non-redundant set used for subsequent analysis.

For Fig. 1a, expression scores for every genomic base pair position were assigned on the basis of how many sequence reads covered each position. The $\log_2$ of the score for each base pair position was then plotted using R/Bioconductor. The numbers of sequence reads drop towards the 5′-end of long genes. To ensure that expression scores are not biased against long genes, scores were determined first by taking the sum of the sequencing expression scores for only 300 bp at the 3′-end of each coding region, or for the entire length if the coding region was <300 bp, and then dividing by the corresponding length used.

Expression-chip hybridization and processing. Total RNA was isolated as described[14], and 0.3 μg RNA were labelled using the standard Affymetrix Genechip eukaryotic hybridization protocols. Hybridizations were performed on Affymetrix Yeast 2.0 Genechip arrays. Scanning was performed on a Genechip Scanner 3000, and data extraction was carried out using Affymetrix GCOS 1.4 (Figs 1a and 4c).

Tiling-chip labelling, hybridization and normalization. Total RNA was isolated as described[14]. Labelling and hybridization to the Affymetrix GeneChip *S. pombe* Tiling 1.0FR arrays were performed as described[5]. Affymetrix CEL files were normalized using the 'normalizeByReference' function from the Bioconductor package 'tilingArray' (http://www.bioconductor.org)[29]. In this procedure, the individual hybridization behaviour of every probe was corrected using the signal of three genomic DNA hybridizations. Genomic DNA was extracted, labelled and hybridized to the Affymetrix GeneChip *S. pombe* Tiling 1.0FR arrays as described[5]. A second normalization step was applied using the signals of intergenic probes as a reference. Finally, between-array normalization and variance-stabilizing transformation were applied using the Bioconductor package 'vsn'.

Pol II ChIP–chip analysis. Chromatin immunoprecipition (ChIP) was performed as described[12] using an antibody specific for the Pol II C-terminal domain (4H8, Abcam). The immunoprecipitated material and input control were amplified in two steps as described[31]. During the second step, dUTPs were added to the PCR mix for subsequent fragmentation of the products. Fragmentation and labelling of the amplified products were performed using the GeneChip WT double-stranded DNA terminal labelling kit (Affymetrix). The duplicated immunoprecipitated samples and corresponding input material were hybridized on four separate Affymetrix GeneChip *S. pombe* Tiling 1.0FR arrays. The $\log_2$ signals of the probes on the input arrays were subtracted from the $\log_2$ signals of the Pol II arrays. The two normalized Pol II data sets were averaged and smoothed using a five-probe moving average. Average gene profiles were created using R and Bioconductor.

Data visualization along genomic coordinates. The tiling-chip data were visualized using the 'plotAlongChrom' function[5] (Figs 2b and 3d). The sequence data were visualized using an in-house R script (Figs 2a and 3a). Normalized sequence scores were generated by dividing the sequencing expression score for a given base pair position by the sum of the expression scores for this base pair position in each condition sequenced.

Novel transcript analysis using tiling-chip data. The normalized data were smoothed using a five-probe moving average. Signal breakpoints in the probe signals along genomic coordinates were then determined using a dynamic programming algorithm for finding a globally optimal fit of a piecewise constant expression profile along genomic coordinates[29]. Segments ≥100 bp and a median probe signal higher than the 75th percentile of the chip and outside of any annotation were selected for visual analysis. To screen for anti-sense transcripts, similar criteria were applied except that the segments had to overlap annotated genes on the opposite strand (Supplementary Table 3).

Novel transcript analysis using sequence data. Stretches of contiguous expression in intergenic regions were identified after removing all UTRs (see below) from the intergenic search space. Novel transcribed regions were required to have a length of ≥70 bp and an average sequence-expression score of ≥5 reads per bp. All predicted novel transcripts were then visually validated to remove inaccurate UTRs before a final manual curation (Supplementary Table 3).

Expression profiling analysis of the novel transcripts. Expression profiles of the novel transcribed regions determined by sequencing and tiling chips were visually inspected from their expression across the 12 biological conditions tested (Supplementary Table 3). For the clustering analysis of Fig. 2c, a Wilcoxon rank sum test was used to determine if the probe signals in each new transcribed region were significantly greater than the signals of a reference set containing probes located outside of any annotated regions in any condition. An expression score was defined as $-\log_2$ of the *P*-value of this test.

UTR determination using tiling-chip data. CEL files were processed as for novel transcripts. The UTR boundaries were the closest breakpoint to the start of an annotated gene, where the median of the four probes immediately upstream of the breakpoint was lower than the one of the four probes downstream of the breakpoint. If no breakpoint could be defined that way and a breakpoint was present <50 bases inside the coding region, the UTR was set to 1. UTRs called inside neighbouring genes or sharing UTR boundaries with neighbouring genes were discarded. UTRs >1,000 nucleotides were discarded, because they were highly enriched in wrong calls based on visual inspection of the data (Supplementary Table 4)

UTR determination using sequence data. UTR lengths were determined by screening for a break in the transcribed region around genes, denoted by positions with sequence scores of 0 or 1, starting from either end of every gene. If a score of 0 was not found in the section between the start and/or end of the neighbouring regions, 1 was used as a cutoff. If no break was found using either cutoff, the UTR was denoted as undetermined (Supplementary Table 4).

Alternate 5′- and 3′-end analysis using tiling-chip data. Genes with UTRs containing several breakpoints caused by 'steps' in the decreasing probe signals moving away from the gene boundaries were automatically selected from 12 biological conditions. A Wilcoxon rank sum test was then used to determine if the probe signals in each region were significantly greater than the signals of a reference set containing probes located outside of any annotated regions in any condition. A score was defined as $-\log_2$ of the *P*-value of this test. Candidate regions with scores >10 in ≥12 conditions were selected for visual inspection (Supplementary Table 6).

Splicing analysis using sequence data. The initial BLAT results generated a set of sequence reads with gaps in the reference sequence (that is, representing potential spliced reads). Spurious matches within this data set caused by poly(A/T) tracts splitting reads between two distant regions in the genome were filtered out using a limit of ≤1 kb for the maximum sequence spanned by trans-reads. The remaining trans-reads were compared to all known and predicted introns for intron validation. Trans-reads that did not span any known introns were clustered on the basis of their splice junctions, where putative junctions had to overlap ±1 bp to belong to the same cluster. Clusters were ranked by the number of novel trans-reads in each cluster and a conservative set of 33,466 reads with ≥6 reads per cluster (defining 485 potential splice sites) were manually curated. 'False-positive' trans-read clusters that did not seem to reflect splicing were mostly within complex repeated regions, and some may reflect errors in the original genome sequence.

Regulated splicing was determined by calculating a ratio of reads that span exon–exon junctions (EE) to those that span the two corresponding exon–intron

junctions (2EI) (Supplementary Table 7). The latter were divided by two to normalize for relative frequency. To obtain a conservative estimate of regulated splicing, the EE:EI ratio for one condition had to be ≥5-times greater than the EE:EI ratio of another stage. Junctions covered by <2 sequence reads in any condition were not considered. Genes that were ≥5-times higher spliced in any meiotic-differentiation stage (M1 to M5) compared to rapidly proliferating cells as well as those that were ≥5-times higher spliced in rapidly proliferating cells compared to ≥1 meiotic-differentiation stage were determined (Fig. 4a). Additional analysis was also performed using absolute read numbers, in cases where ratios could not be calculated because of 0 values. In these cases, to obtain a conservative estimate of regulated splicing, where EE = 0 in rapidly proliferating cells, the EE in ≥1 meiotic-differentiation stage was required to be >6. With EE = 1 or 2 but EI = 0 in rapidly proliferating cells, the EE in ≥1 meiotic-differentiation stage was required to be 8 or 9, respectively. With EE ≥3 in rapidly proliferating cells, the EE in ≥1 meiotic-differentiation stage was required to be ≥5-times higher than in rapidly proliferating cells, or ≥20-times higher when identifying introns spliced more efficiently in rapidly proliferating cells to account for the greater sequence depth in this condition (Supplementary Table 8).

Measurement of splicing efficiency by quantitative RT–PCR. To test the relationship between transcription rate and splicing efficiency (Fig. 4d and Supplementary Fig. 7), the uracil-inducible *urg1* promoter was integrated upstream of *spo6* and *spn7* (ref. 30). Cells were grown in exponential phase for 16 h in minimal medium (MM) in the absence of uracil. A cell sample was then harvested, and uracil was added to the remaining culture at a final concentration of 2 mg ml^{-1}. Further cell samples were harvested 15 min and 60 min after uracil addition.

spo6 and *spn7* are putative targets of Mei4 and were induced in a strain over-expressing Mei4 under the control of the *nmt1* promoter[27]. Such a strain (Supplementary Table 1) was grown in the presence of thiamine to early exponential phase. A cell sample was then harvested before the cells were diluted and was grown for 18 h in the absence of thiamine.

Primers were designed inside the exons 1 and 2 of *spo6* and across the exon 1/intron 1 and exon 2/intron 2 junctions. Similarly, primers were designed inside exons 1 and 4 of *spn7* and across the exon1/intron1 and intron 3/exon 4 junctions. RNA was extracted and qRT–PCR performed as described[30]. The data were normalized to the signal of the *fba1* control gene. No signals above background levels were detected in control runs in the absence of reverse transcriptase.

Curation methods. Novel transcribed regions were converted to gff3 format and visualized in the context of the existing annotation using Artemis software and methods described previously[13]. The corresponding sequence plots were examined and discrete features designated 'non-coding RNAs'. Manual inspection of the strand-specific tiling-chip data identified several 'antisense' transcripts. 'Non-coding RNAs' were inspected for the presence of methionine-containing ORFs >60 amino acids, identifying three protein-coding genes. Less discrete features, which may correspond to transcriptional noise, occurring mainly in low-complexity regions were designated 'miscellaneous features'. Some transcribed features were clearly related to their proximal genes and curated as 5′ and 3′ UTRs (occasionally intron-containing).

Sequence trans-reads obtained from proliferating cells validated 3,796 of the 4,811 known and predicted introns, and trans-reads only obtained from meiotic cells validated an additional 666 introns (Supplementary Table 7). The remaining 349 introns either were in poorly expressed genes with insufficient sequence reads, or were not spliced under any of the conditions tested. Among the latter, manual inspection coupled with homology searches and intron branch, acceptor and donor consensus-sequence data allowed refinement of 25 protein-coding gene structures, and deletion of 6 unsupported intron-containing genes. A number of the introns confirmed by trans-read sequences were not previously annotated in the database. These 'false negative' introns were mapped onto the genomic sequence and used to identify 22 new genes and revise a further ~60 gene structures.

All these alterations have been incorporated in *S. pombe* gene database (http://www.genedb.org/genedb/pombe/). The new transcribed regions are listed in Supplementary Table 3, and the corrected gene structures are listed at http://www.genedb.org/genedb/pombe/coordChanges.jsp.

31. Bernstein, B. E., Humphrey, E. L., Liu, C. L. & Schreiber, S. L. The use of chromatin immunoprecipitation assays in genome-wide analyses of histone modifications. *Methods Enzymol.* **376**, 349–360 (2004).

Vol 453|26 June 2008|doi:10.1038/nature07003

LETTERS

Analysis of a spatial orientation memory in *Drosophila*

Kirsa Neuser[1]†, Tilman Triphan[1], Markus Mronz[1], Burkhard Poeck[1] & Roland Strauss[1]†

Flexible goal-driven orientation requires that the position of a target be stored, especially in case the target moves out of sight. The capability to retain, recall and integrate such positional information into guiding behaviour has been summarized under the term spatial working memory[1]. This kind of memory contains specific details of the presence that are not necessarily part of a long-term memory. Neurophysiological studies in primates[2] indicate that sustained activity of neurons encodes the sensory information even though the object is no longer present. Furthermore they suggest that dopamine transmits the respective input to the prefrontal cortex, and simultaneous suppression by GABA spatially restricts this neuronal activity[3]. Here we show that *Drosophila melanogaster* possesses a similar spatial memory during locomotion. Using a new detour setup, we show that flies can remember the position of an object for several seconds after it has been removed from their environment. In this setup, flies are temporarily lured away from the direction towards their hidden target, yet they are thereafter able to aim for their former target. Furthermore, we find that the GABAergic (stainable with antibodies against GABA) ring neurons[4] of the ellipsoid body in the central brain are necessary and their plasticity is sufficient for a functional spatial orientation memory in flies. We also find that the protein kinase S6KII (*ignorant*)[5] is required in a distinct subset of ring neurons to display this memory. Conditional expression of S6KII in these neurons only in adults can restore the loss of the orientation memory of the *ignorant* mutant. The S6KII signalling pathway therefore seems to be acutely required in the ring neurons for spatial orientation memory in flies.

Previous studies have shown that walking flies heading for an object maintain their direction even when the target disappears[6]. This persistence of orientation can last for several seconds, indicating that flies store the position of, or the path towards, the hidden object for further targeting. We therefore proposed that flies form a spatial memory for objects that is similar to the working memory in vertebrates. To investigate this putative memory in *Drosophila* we established a detour paradigm for walking flies (Fig. 1a–f). Single flies were

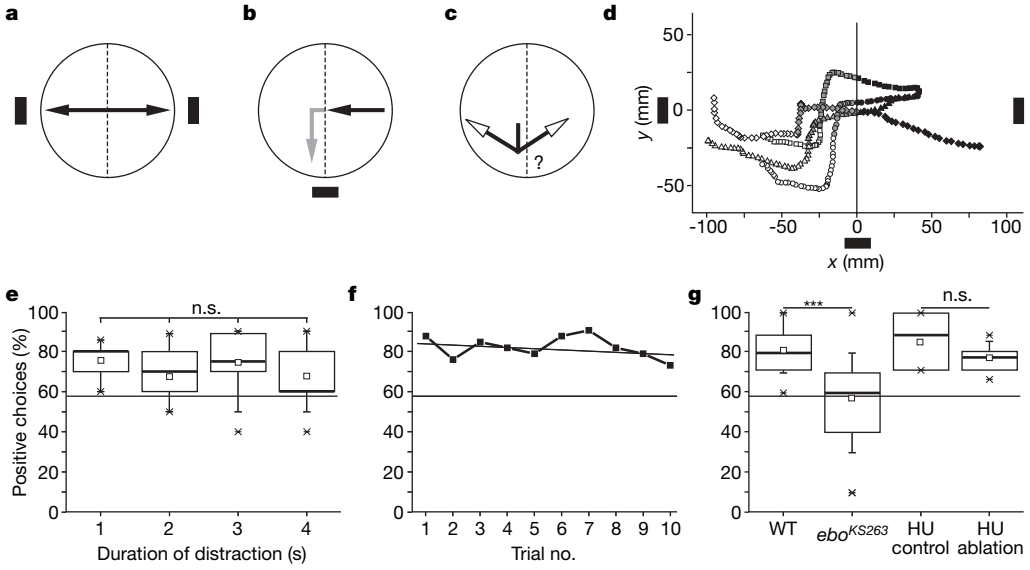


Figure 1 | Orientation memory in the detour paradigm. **a**, A fly patrols between two vertical stripes shown on a cylindrical screen. On crossing the midline, the stripes disappear and simultaneously another one appears laterally to the fly. **b**, After the fly has turned towards the distracter for 1 s this stripe also disappears. **c**, Subsequently, it is determined whether or not the fly turns back towards its original target. **d**, Walking traces of four wild-type males. Black symbols, initial phase (**a**); grey symbols, distractor phase (**b**); open symbols, memory phase (**c**). The maximum duration of a trace is 20 s. **e**, Prolonged distraction does not change the orientation memory ($P = 0.22$). Bold horizontal lines represent the medians, squares the means, boxes the 25% and 75% quartiles, whiskers the 10% and 90% quantiles, and stars the extreme values. **f**, Percentage of positive choices for each of ten consecutive trials ($n = 73$; $r^2 = 0.08$). **g**, Mushroom-body ablation with hydroxyurea (HU) does not impair the memory in comparison with mock treatment ($P = 0.25$). In contrast, ebo^{KS263} mutants show a reduced performance compared to wild-type (WT) flies ($P < 10^{-3}$), as indicated by three asterisks. n.s., not significant. The horizontal line in **e**–**g** indicates the 58% chance level.

[1]Lehrstuhl für Genetik und Neurobiologie, Biozentrum, Universität Würzburg, Am Hubland, D-97074 Würzburg, Germany. †Present address: Institut für Zoologie III – Neurobiologie, Universität Mainz, Col.-Kleinmann-Weg 2, D-55099 Mainz, Germany.

put into a cylindrical virtual-reality arena[7], in which two dark vertical stripes were presented at opposite sides (Fig. 1a). Normally, flies patrol between the two visual objects for a considerable length of time[8]. In our new paradigm, the stripes disappeared when the fly crossed the invisible midline of the circular walking platform, and a new target appeared laterally at a 90° angle to the fly. In most cases wild-type flies turned towards this new target if it was presented for more than 500 ms (Fig. 1b). After the fly had oriented itself towards the new object (deviation of the fly's longitudinal body axis from the ideal course to the stripe below ±15°), this target also disappeared within 1 s and no objects were visible to the fly. We then determined whether the fly turned back to continue its approach to its initial, but still invisible, target (Fig. 1c–f; see Methods and Supplementary Video). The walking traces (Fig. 1d) reveal a direct course towards the former location of the first target. The flies therefore retained positional information on the former object, although it was no longer present in the environment.

Wild-type (Canton-S) flies recall the old target and integrate it into a guided behaviour with a median frequency of 80% (Fig. 1e) as measured in ten consecutive trials for each fly. Longer presentation of the distracter stripe did not significantly change the percentage of positive choices (Fig. 1e). These data strongly suggest that flies stored the relative position of the first target in a spatial orientation memory for at least 4 s. To exclude the possibility that flies used chemical traces of former runs for their orientation we randomly changed the absolute positions of the stripes after each trial. As a result of this randomization, flies had to update their memory continuously. Moreover, we could not observe any training effect, because the frequency of positive turns did not change during the ten consecutive trials (Fig. 1f). Similar performances were observed when two opposing distracters were presented to the fly (data not shown). We consider this orientation memory for vanished objects to be idiothetic. Because no visible landmarks were presented to the fly after the distracter disappeared, the fly could not use a stored reference picture of the environment for its guidance. We therefore suggest that the fly uses online stored information of its own angle towards the former target, a strategy known as path integration. Path integration has been shown to be used by other insects, such as ants and bees, to navigate through a familiar landscape[9].

In an attempt to localize this type of memory to discrete parts of the insect brain, several mutant lines with structural central-complex defects of *Drosophila* were analysed[10]. The central complex is composed of four different neuropils (Fig. 2g) and has been implicated in supervising motor output during locomotion[10,11]. First tests showed that the persistence of orientation towards a removed target is reduced or lost whenever the ellipsoid body of the central complex was defective (Supplementary Fig. 1). We therefore tested the *ellipsoid body open* mutant (ebo^{KS263})[10] in the detour paradigm; these flies did not show a preference for the first target after the detour, suggesting that an intact ellipsoid body is required for establishing a spatial orientation memory. In contrast, the use of hydroxyurea to ablate the mushroom bodies, which are important in olfactory memory[12], did not disturb the orientation memory (Fig. 1g).

One prominent type of neuronal cells of the ellipsoid body is the group of GABAergic ring neurons[4]. The fibres of these neurons run in a prominent tract, the RF tract (ring-neuron and tangential fan-shaped-body neuron tract), and form bushy thin endings in the ipsilateral lateral triangle and bleb-like endings in the ellipsoid body (Fig. 2g). Four different kinds of ring neuron (R1–R4) can be distinguished by their arborization pattern around the ellipsoid body canal. R1–R3 neurons project outwards from the ellipsoid body canal, whereas the arborization of R1 is restricted to the inner zone (Fig. 2f), that of R2 to the outer zone, and that of R3 to both zones (Fig. 2e). R4 neurons project from the periphery inwards and arborize in the outermost zone (Fig. 2d)[13]. We next proposed that the ring neurons might be necessary for the orientation memory. We used the GAL4/UAS system[14] to silence distinct subsets of ring neurons through the expression of tetanus toxin (TNT)[15] by using the GAL4 driver lines c232, c481 and c105 (ref. 13) (Fig. 2d–f and Supplementary Fig. 2). For temporal control, we induced TNT conditionally by using the temperature-sensitive GAL4 repressor $GAL80^{ts}$ under the control of the ubiquitous *Tubulin* promoter

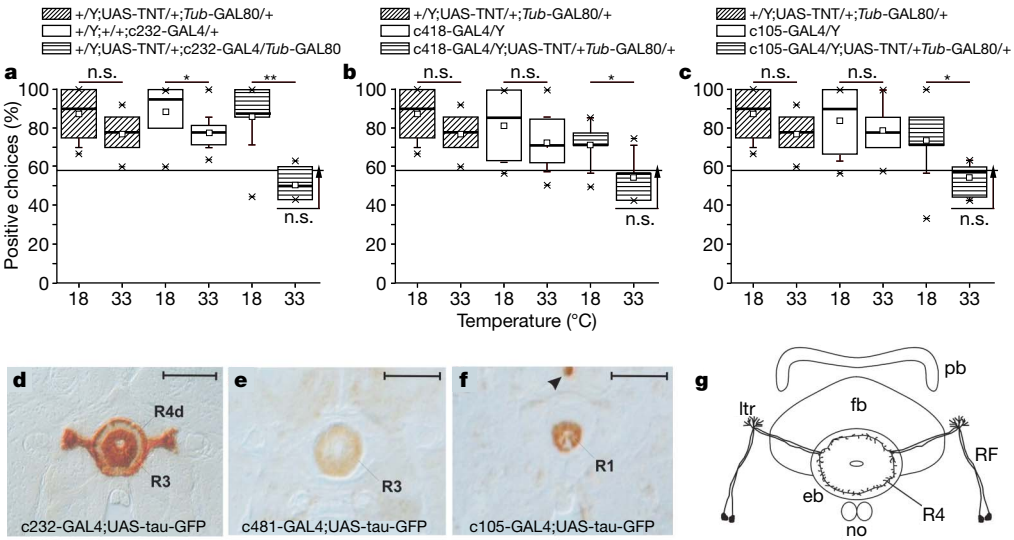


Figure 2 | The ellipsoid-body ring neurons are necessary for orientation memory. **a–c**, Conditional induction of tetanus toxin in R4 and R3 ring neurons (driver line c232) (**a**), R3 ring neurons (line c481) (**b**) and R1 ring neurons (line c105) (**c**) leads to total memory loss (pairwise comparisons, non-induced (18 °C) versus induced (33 °C), $P = 0.004$, $P = 0.01$ and $P = 0.05$). All experimental groups chose randomly ($P = 0.18$, $P = 0.43$ and $P = 0.15$), whereas all control groups preferred the former target (full statistical account in Supplementary Table 1). Box–whisker symbols are as described in the legend to Fig. 1. Asterisk, $P < 0.05$; two asterisks, $P < 0.01$; **d–f**, Frontal paraffin sections of adult fly heads showing the expression patterns of the driver lines (UAS-tau-GFP reporter; scale bars 50 μm). **d**, Line c232, showing strong staining of distal R4 and R3 ring neurons. **e**, Line c481, showing exclusive expression in R3 ring neurons. **f**, Line c105, showing R1 ring neurons and the ocellar nerve (arrowhead). **g**, Frontal drawing of the *Drosophila* central complex (pb, protocerebral bridge; fb, fan-shaped body; eb, ellipsoid body; no, noduli). The location of ring-neuron type R4 is shown. Ring neurons have their perikarya ventral to the ellipsoid body. Their axons project by way of the RF tract to the lateral triangles (ltr), where they form spiny arborization, and on to the ellipsoid body.

(*Tub*-GAL80^{ts})[16]. Experimental and control flies were raised at 18 °C, tested within the detour paradigm, and retested after the induction of TNT. Pairwise comparison revealed that the preference for the original target was lost whenever the toxin was expressed in ring neurons of the ellipsoid body (Fig. 2). This finding confirms our hypothesis that the ellipsoid-body ring neurons are necessary components of the orientation memory.

To investigate which molecular pathways are involved in this kind of memory, we first focused on the cyclic-AMP signalling pathway. Variable levels of cAMP have been shown to have a crucial function in memory formation during associative learning in *Drosophila*[17,18]. cAMP levels are modulated by the opposing actions of adenylyl cyclases and cAMP phosphodiesterases. Mutants for the adenylyl cyclase gene *rutabaga* (*rut*^{1} and *rut*^{2080}) were unable to target visual objects and could not be tested in our paradigm. We therefore tested mutants of the *dunce* gene (*dnc*), which encodes a cAMP phosphodiesterase, in the detour paradigm. The *dnc*^{1} mutant is a hypomorph and displays about half of the enzyme activity in the wild type[19]. *dnc*^{1} mutant flies show deficits in several paradigms of associative classical learning[20] and operant conditioning[21]. In contrast, *dnc*^{1} mutants showed no defects in the detour paradigm (Fig. 3a), indicating that a tight modulation of cAMP levels might not be critically required for spatial orientation memory.

Another molecule involved in memory formation in *Drosophila* is a member of the ribosomal serine kinase family. *ignorant* (*ign*) encodes the S6 kinase II (S6KII), which interacts with mitogen-activated protein (MAP) kinase signalling in *Drosophila*[22] and vertebrates[23]. S6KII does not seem to be involved in cAMP signalling pathways. The null allele *ign*$^{58/1}$ has been shown to be defective in classical aversive conditioning and operant learning[5]. We therefore tested *ign*$^{58/1}$ flies in the detour paradigm. Although the mutants readily targeted visible objects, they showed no directional preference for the position of the original target after it disappeared, suggesting that they had lost their memory (Fig. 3a). In contrast, walking speed, walking activity and orientation towards visual objects were similar to those of the wild type (Supplementary Fig. 3). Next we examined whether *ign* is required in the ring neurons targeted by c232-GAL4 with the use of a UAS-*ign* RNA-mediated interference (RNAi) effector line[24]. RNAi silencing in these ring neurons decreased the performance by half. This decrease in memory constitutes only a partial phenocopy of the null mutant, because the performance was not significantly different from that of the wild type or *ign*$^{58/1}$ (Fig. 3b). Nevertheless, we interpret this result to suggest that *ign* is required in the ring neurons for spatial orientation memory.

To address the question of whether restoring S6KII levels is sufficient for regaining memory, we performed neuron-specific rescue experiments in the *ign*$^{58/1}$ mutant background. S6KII was expressed pan-neuronally with *Appl*-GAL4 (ref. 25) and *elav*-GAL4 (ref. 26),

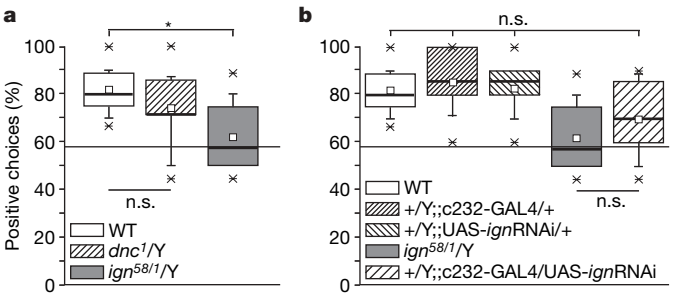


Figure 3 | S6KII activity is necessary for orientation memory. **a**, *dnc*^{1} mutant flies do not differ from wild-type (WT) flies in their memory performance ($P = 0.34$), whereas *ign*$^{58/1}$ mutant flies show a complete memory loss (asterisk; $P = 0.012$). **b**, *ign*-RNAi silencing in the R3 and R4 ring neurons leads to a partial decrease in orientation memory (RNAi versus *ign*$^{58/1}$, $P = 0.27$; WT versus RNAi, $P = 0.09$). Box–whisker symbols are as described in the legend to Fig. 1.

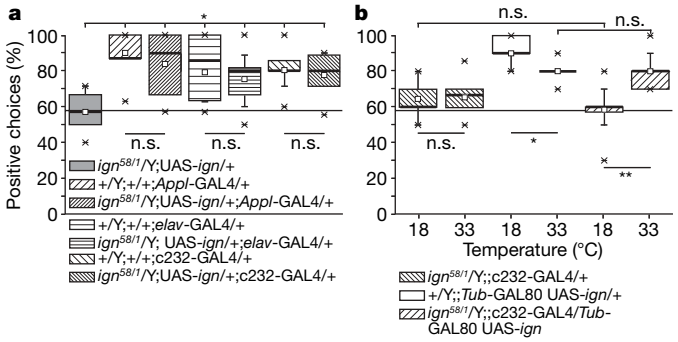


Figure 4 | S6KII activity in the ring neurons is sufficient to restore orientation memory. **a**, The memory loss of the *ign*$^{58/1}$; UAS-*ign* mutant ($P = 0.04$) can be rescued by expressing S6KII protein either pan-neuronally (*Appl*-GAL4, $P = 0.58$; *elav*-GAL4, $P = 0.63$) or exclusively in R3 and R4 ring neurons (c232-GAL4, $P = 0.83$). **b**, Conditional expression of *ign* in the R3 and R4 ring neurons is sufficient to restore the memory in the adult (two asterisks; $P = 0.001$). Asterisk, $P < 0.05$. Box–whisker symbols are as described in the legend to Fig. 1.

and also specifically in the R3 and R4 ring neurons with c232-GAL4. In all three cases we observed a complete rescue (Fig. 4a). Next we examined whether *ign* function in the R3 and R4 ring neurons is acutely required for orientation memory. We therefore again made use of the GAL80^{ts} transgene to rescue the *ign* phenotype only in the adult. Conditional expression of S6KII only in the R3 and R4 ring neurons resulted in a perfect rescue of the *ign* mutant (Fig. 4b). Our result—that acute S6KII expression in the R3 and R4 ring neurons accomplished a complete rescue—reveals that this very narrow subset of cells is sufficient for regaining a functional orientation memory. It has been reported that *Drosophila* S6KII negatively regulates extracellular signal-regulated kinases (ERKs) by acting as a cytoplasmic anchor of the MAP kinase[22]. Further studies will determine whether the MAP kinase signalling pathway is required for this kind of memory task.

The relevant ring neurons use the inhibitory neurotransmitter GABA. Their circuitry and interconnections within the ellipsoid body are not yet known. Expression of the dDA1 dopamine receptor in the ellipsoid body has recently been shown[27]. It is therefore possible that the same neurotransmitter systems as those used for visual–spatial memory in the monkey prefrontal cortex[3] are used to establish orientation memory in the central complex of flies.

METHODS SUMMARY

Fly strains. Fly strains were raised on standard medium under a 14 h/10 h light/dark cycle. GAL4 strains c232, c481 and c105 were provided by D. Armstrong, and the UAS-*ign* and the *ign*$^{58/1}$ strains by M. Heisenberg. *dnc*^{1}, *Appl*-GAL4, *elav*-GAL4, UAS-TNT, *Tub*-GAL80^{ts} and UAS-tau–GFP stocks were obtained from the Bloomington stock centre, and the UAS-*ign*-RNAi line from the Vienna Drosophila RNAi Center. Wild-type Canton-S flies served as control.

Statistics. Ten approaches of at least ten flies per genotype were recorded; the percentage of choices towards the vanished object was calculated for each fly, and the median frequency of positive choices was determined. Random behaviour would result in a preference value of 58% because there is a greater probability that a right turn will follow a left turn, and vice versa[28]. Box–whisker plots show the median (bold line), the mean (square), 25% and 75% quartiles (box), 10% and 90% quantiles (whiskers) and extreme values (stars). Because some data were not normally distributed (Shapiro–Wilks *W*-test), we used the Kruskal–Wallis analysis of variance test for multiple comparisons, the Mann–Whitney *U*-test for independent comparisons and the sign test for dependent pairwise comparisons. The one-sample sign test was used to compare groups with the random value. Statistical analyses were performed with STATISTICA 7.0 (α level 0.05 in all cases). See Supplementary Table 1 for all statistical calculations.

Histology. Ablations of the mushroom bodies with hydroxyurea were performed as described[12]. Completeness of ablation was assessed with paraffin histology[29] for each test fly. Expression analysis of the GAL4 driver lines with a monoclonal anti-bovine TAU antibody (1:200 dilution; Sigma) on paraffin sections was conducted as described[30].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 18 February; accepted 15 April 2008.
Published online 28 May 2008.

1. Postle, B. R. Working memory as an emergent property of the mind and brain. *Neuroscience* **139,** 23–38 (2006).
2. Chafee, M. V. & Goldman-Rakic, P. S. Matching patterns of activity in primate prefrontal area 8a and parietal area 7ip neurons during a spatial working memory task. *J. Neurophysiol.* **79,** 2919–2940 (1998).
3. Williams, G. V. & Castner, S. A. Prefrontal cortex and working memory processes. *Neuroscience* **139,** 251–261 (2006).
4. Hanesch, U., Fischbach, K.-F. & Heisenberg, M. Neuronal architecture of the central complex in *Drosophila melanogaster*. *Cell Tissue Res.* **257,** 343–366 (1998).
5. Putz, G., Bertolucci, F., Raabe, T., Zars, T. & Heisenberg, M. The S6KII (*rsk*) gene of *Drosophila melanogaster* differentially affects an operant and a classical learning task. *J. Neurosci.* **24,** 9745–9751 (2004).
6. Strauss, R. & Pichler, J. Persistence of orientation toward a temporarily invisible landmark in *Drosophila melanogaster*. *J. Comp. Physiol. A* **182,** 411–423 (1998).
7. Strauss, R., Schuster, S. & Götz, K. G. Processing of artificial visual feedback in the walking fruit fly *Drosophila melanogaster*. *J. Exp. Biol.* **200,** 1281–1296 (1997).
8. Bülthoff, H., Götz, K. G. & Herre, M. Recurrent inversion of visual orientation in the walking fly, *Drosophila melanogaster*. *J. Comp. Physiol. A* **148,** 471–481 (1982).
9. Collett, T. S. & Collett, M. Memory use in insect visual navigation. *Nature Rev. Neurosci.* **3,** 542–552 (2002).
10. Strauss, R. & Heisenberg, M. A higher control center of locomotor behavior in the *Drosophila* brain. *J. Neurosci.* **13,** 1852–1861 (1993).
11. Strauss, R. The central complex and the genetic dissection of locomotor behaviour. *Curr. Opin. Neurobiol.* **12,** 633–638 (2002).
12. De Belle, J. S. & Heisenberg, M. Associative odor learning in *Drosophila* abolished by chemical ablation of mushroom bodies. *Science* **263,** 692–695 (1994).
13. Renn, S. C. *et al.* Genetic analysis of the *Drosophila* ellipsoid body neuropil: organization and development of the central complex. *J. Neurobiol.* **41,** 189–207 (1999).
14. Brand, A. H. & Perrimon, N. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development* **118,** 401–415 (1993).
15. Sweeney, S. T., Broadie, K., Keane, J., Niemann, H. & ÒKane, C. J. Targeted expression of tetanus toxin light chain in *Drosophila* specifically eliminates synaptic transmission and causes behavioral defects. *Neuron* **14,** 341–351 (1995).
16. McGuire, S. E., Le, P. T., Osborn, A. J., Matsumoto, K. & Davis, R. L. Spatiotemporal rescue of memory dysfunction in *Drosophila*. *Science* **302,** 1765–1768 (2003).
17. McGuire, S. E., Deshazer, M. & Davis, R. L. Thirty years of olfactory learning and memory research in *Drosophila melanogaster*. *Prog. Neurobiol.* **76,** 328–347 (2005).
18. Liu, G. *et al.* Distinct memory traces for two visual features in the *Drosophila* brain. *Nature* **439,** 551–556 (2006).
19. Davis, R. L. & Kiger, J. A. Jr. *Dunce* mutants of *Drosophila melanogaster*: mutants defective in the cyclic AMP phosphodiesterase enzyme system. *J. Cell Biol.* **90,** 101–107 (1981).
20. Dudai, Y., Jan, Y. N., Byers, D., Quinn, W. G. & Benzer, S. *dunce*, a mutant of *Drosophila* deficient in learning. *Proc. Natl Acad. Sci. USA* **73,** 1684–1688 (1976).
21. Wustmann, G., Rein, K., Wolf, R. & Heisenberg, M. A new paradigm for operant conditioning of *Drosophila melanogaster*. *J. Comp. Physiol. A* **179,** 429–436 (1996).
22. Kim, M. *et al.* Inhibition of ERK-MAP kinase signaling by RSK during *Drosophila* development. *EMBO J.* **25,** 3056–3067 (2006).
23. Myers, A. P., Corson, L. B., Rossant, J. & Baker, J. C. Characterization of mouse Rsk4 as an inhibitor of fibroblast growth factor-RAS-extracellular signal-regulated kinase signaling. *Mol. Cell. Biol.* **24,** 4255–4266 (2004).
24. Dietzl, G. *et al.* A genome-wide transgenic RNAi library for conditional gene inactivation in *Drosophila*. *Nature* **448,** 151–156 (2007).
25. Torroja, L., Chu, H., Kotovsky, I. & White, K. Neuronal overexpression of APPL, the *Drosophila* homologue of the amyloid precursor protein (APP), disrupts axonal transport. *Curr. Biol.* **9,** 489–492 (1999).
26. Luo, L., Liao, Y. J., Jan, L. Y. & Jan, Y. N. Distinct morphogenetic functions of similar small GTPases: *Drosophila Drac1* is involved in axonal outgrowth and myoblast fusion. *Genes Dev.* **8,** 1787–1802 (1994).
27. Kim, Y. C., Lee, H. G., Seong, C. S. & Han, K. A. Expression of a D1 dopamine receptor dDA1/DmDOP1 in the central nervous system of *Drosophila melanogaster*. *Gene Expr. Patterns* **3,** 237–245 (2003).
28. Mronz, M. *Die visuell motivierte Objektwahl laufender Taufliegen (*Drosophila melanogaster*)–Verhaltensphysiologie, Modellbildung und Implementierung in einem Roboter.* PhD thesis, Univ. Würzburg (2004).
29. Heisenberg, M. & Boehl, K. Isolation of anatomical brain mutants of *Drosophila* by histological means. *Z. Naturforsch. C* **34,** 143–147 (1979).
30. Botella, J. A. *et al.* Deregulation of the Egfr/Ras signaling pathway induces age-related brain degeneration in the *Drosophila* mutant *vap*. *Mol. Biol. Cell* **14,** 241–250 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Heisenberg for discussions and continuous support, D. Kretzschmar for reading the manuscript, and E. Stepien-Bötsch for experimental contributions. This work was supported by the Deutsche Forschungsgemeinschaft (SFB 554-B7, GRK 1156).

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.S. (rstrauss@uni-mainz.de).

METHODS

Detour paradigm. In all experiments 3–5-day-old male flies were used. One day before the experiment their wings were shortened under cold anaesthesia. Flies were kept individually overnight on filter paper soaked with pure water. Experiments were performed in a cylindrical panorama composed of 5760 light-emitting diodes on a walking platform 160 mm in diameter, surrounded by a water-filled moat to prevent flies from escaping. Video-camera pictures at 20 Hz taken from above were processed on a PC and reduced to 5 Hz for noise reduction with a computer program[10] modified by T.T.

At the beginning of an experiment the virtual-reality arena showed two dark opposing stripes with 12° horizontal and 50° vertical viewing angles as seen from the centre. An experiment started by placing a single fly into the centre of the arena. As soon as the fly had travelled once between these two stripes and crossed the midline again, the stripes disappeared. Simultaneously a vertical distracter stripe appeared laterally to the fly. Whenever the orientation of the fly's actual path increment towards this distracting stripe deviated by less than ±15° from the direct path to the distracter, it disappeared after 1 s and no stripes were visible thereafter. We used the 1-s presentation of the distracter in all experiments except for those in Fig. 1e. Linking the time of disappearance to the ±15° criterion guarantees that the fly actually fixated on that new object. Subsequently it was determined whether the fly turned back to its previous, still invisible, target. To detect such a turn the computer program first calculated the average angular deviation from the direct path to the distracter from the fly's three latest path increments before the distracter disappeared. By definition, a choice was made as soon as a subsequent path increment deviated from this latter mean by more than ±45°, either towards (termed a positive choice) or away from (negative choice) the initial target stripe. After such a decision had been detected, the two opposing stripes appeared again at a new, randomized position and the next of a total of ten consecutive trials started whenever the fly crossed the midline again.

Conditional expression via the GAL80^{ts} system. For the conditional expression of UAS-TNT (ref. 15) or UAS-*ign* (ref. 5), a double homozygous stock with UAS-TNT; *Tub*-GAL80^{ts} (ref. 16) and a recombinant chromosome *Tub*-GAL80^{ts}, UAS-*ign* were established, respectively. Experimental and control flies were reared at 18 °C and 3–5-day-old males, with clipped wings, were tested in the detour paradigm at room temperature (20–22 °C). The individual flies were then incubated overnight at the permissive temperature (33 °C). After a minimum recovery time of 3 h at room temperature the flies were tested again.

LETTERS

Hippocampus-independent phase precession in entorhinal grid cells

Torkel Hafting[1]*, Marianne Fyhn[1]*, Tora Bonnevie[1], May-Britt Moser[1] & Edvard I. Moser[1]

Theta-phase precession in hippocampal place cells[1] is one of the best-studied experimental models of temporal coding in the brain. Theta-phase precession is a change in spike timing in which the place cell fires at progressively earlier phases of the extracellular theta rhythm as the animal crosses the spatially restricted firing field of the neuron[1–5]. Within individual theta cycles, this phase advance results in a compressed replication of the firing sequence of consecutively activated place cells along the animal's trajectory[2,6–8], at a timescale short enough to enable spike-time-dependent plasticity between neurons in different parts of the sequence. The neuronal circuitry required for phase precession has not yet been established. The fact that phase precession can be seen in hippocampal output stuctures such as the prefrontal cortex[9] suggests either that efferent structures inherit the precession from the hippocampus or that it is generated locally in those structures. Here we show that phase precession is expressed independently of the hippocampus in spatially modulated grid cells[10,11] in layer II of medial entorhinal cortex, one synapse upstream of the hippocampus. Phase precession is apparent in nearly all principal cells in layer II but only sparsely in layer III. The precession in layer II is not blocked by inactivation of the hippocampus, suggesting that the phase advance is generated in the grid cell network. The results point to possible mechanisms for grid formation and raise the possibility that hippocampal phase precession is inherited from entorhinal cortex.

To explore the relationship between spatial and temporal oscillations in grid cells[12–14], we recorded electroencephalogram (EEG) and spike times in dorsocaudal medial entorhinal cortex (MEC) while rats shuttled back and forth on a linear track that covered multiple periods of the grid (Figs 1 and 2, and Supplementary Figs 1–5). Running was associated with strong and continuous theta rhythmicity in both MEC and hippocampus (Figs 1b, c and 2b, c, and Supplementary Figs 6 and 7). In MEC, grid cells had multiple distinct firing locations that were stable across laps[11] (Figs 1d–f and 2d–f). The same cells showed clear spatial periodicity in a two-dimensional open field, with grid spacings between 30 and 70 cm and modulation by the rat's running direction in layer III but not in layer II (Figs 1a and 2a)[11,15].

Grid cells showed clear theta-phase precession as animals passed through individual grid fields on the track. The most reliable precession was observed in layer II (Figs 1c, g and 3a, and Supplementary Fig. 2). When the rat entered one of the grid fields, the cell began discharging on the ascending slope of the local theta oscillation, between 180° (trough) and 360° (peak), but nearer the trough than the peak of the theta cycle (Figs 1g and 3a, and Supplementary Figs 2 and 8). The initial phase was 222 ± 62° (circular mean ± angular deviation; 133 fields). As the rat crossed the field, the phase advanced progressively over successive cycles of the theta rhythm (phase at the exit of the field: 59 ± 78°; entry versus exit: $F_{2,132} = 16.8$, $P < 0.001$; Hotelling test for paired samples of angles; Supplementary Figs 8 and 9). The phase advance never exceeded 360°. In many cells the

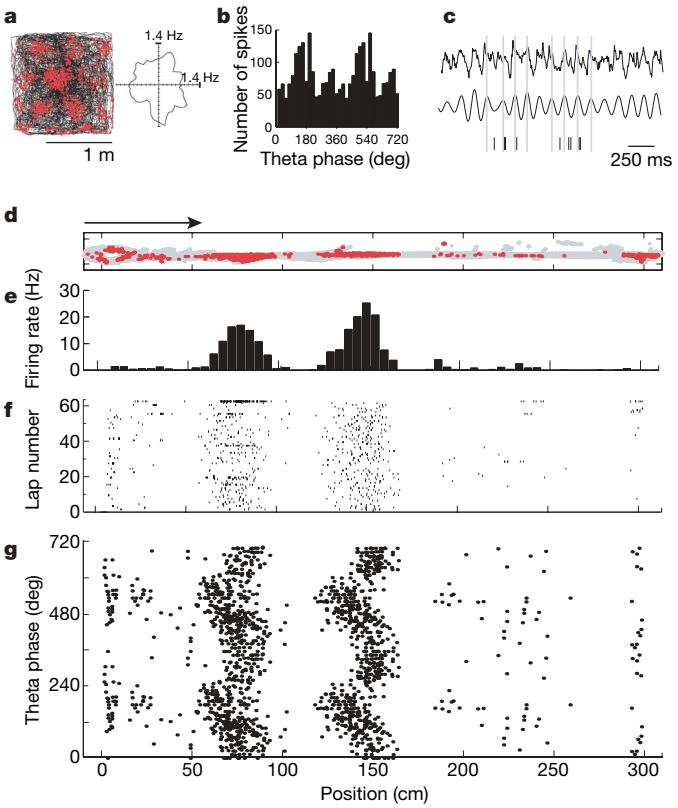


Figure 1 | Phase precession in a layer II cell in MEC. a, Left: firing field of a grid cell during running in an open field (black, trajectory; red, individual spikes). Right: firing rate as a function of running direction (bin size 6°). **b**, Distribution of firing rate within the theta cycle for the cell in **a** during running on a linear track (bins of 20°, two theta cycles, peak of local theta rhythm is 0°). **c**, Local entorhinal EEG with spike times for the layer II cell in **a** and **b** during 2.1 s of track running. Vertical ticks are individual spikes. Top: unfiltered EEG trace. Bottom: filtered at 6–11 Hz. Vertical grey lines indicate 0°. **d**–**g**, Rate distribution and phase relationship for spikes of the cell in **a**–**c** during running from left to right (entire trial). **d**, Trajectory (grey) with locations of individual spikes (red). Arrow indicates running direction. **e**, Linearized spatial firing rate map (bins of 5 cm). **f**, Raster plot indicating spike positions (vertical ticks) on the track. **g**, Theta phase as a function of position (two theta cycles). In **d** and **e** all spikes are shown except at the turning points; in **f** and **g** spikes are shown for all epochs with running at more than 10 cm s^{-1}. Note the gradual advance of firing phase as the rat passes through each field. For additional examples see Supplementary Fig. 2.

[1]Kavli Institute for Systems Neuroscience and Centre for the Biology of Memory, Norwegian University of Science and Technology, NO-7489 Trondheim, Norway.
*These authors contributed equally to this work.

progression was bimodal, as in hippocampal pyramidal cells[2,4,16], with a slow advance during the first 180° and a faster advance and a wider distribution during the second part (Fig. 3a and Supplementary Fig. 2).

The correlation between phase and position at the phase rotation that gave the regression line with the largest explained variance R^2 was negative for 109 of 133 fields (82.0%; $Z = 7.4$, $P < 0.001$, with an expected P value of 0.50; Fig. 3b and Supplementary Fig. 2). Positive correlations were observed only when the explained variance was low (89.2% of the 93 fields with $R^2 > 0.10$ and 100% of the 34 cells with $R^2 > 0.25$ had negative correlation coefficients; Supplementary Fig. 10). The proportion of fields with a negative correlation between phase and position was comparable to that of simultaneously recorded pyramidal cells in the hippocampus (CA3: 86.4%, 43 fields; CA1: 85.2%, 54 fields; both $Z < 1.4$, not significant (n.s.); Fig. 3b and Supplementary Fig. 4). The correlation between phase and position in the layer II cells was $r = -0.29 \pm 0.03$ (mean ± s.e.m.), which also was similar to that in CA3 and CA1 ($r = -0.29 \pm 0.05$). The mean slope of the best-fit regression line was $-2.77 \pm 0.31°\,\mathrm{cm}^{-1}$. The precession rate was faster for cells at the dorsal end of MEC, where grid fields are small, than for cells at more ventral locations, where grid spacing is larger[10,11,15] (Supplementary Fig. 11; distance from dorsal border of MEC versus absolute value of the slope of the regression line: $r = -0.45$, $P < 0.001$; field length versus absolute value of the slope: $r = -0.61$, $P < 0.001$).

Strong phase precession was also observed in a small sample of layer V cells (Supplementary Fig. 12) but not in layer III (Figs 2 and 3). The overall level of phase precession in layer III was much weaker than in layer II. Among the 82 fields that could be confidently localized to layer III, only about 20 (about 25%) showed visible phase precession across the entire theta cycle (Supplementary Fig. 3). In the other fields, the spikes remained confined to the trough of the theta cycle (about 25%), or the cells fired non-differentially with respect to phase (about 50%). The correlation between phase and position was negative for only 45 of the fields (54.9%; $Z = 0.89$, n.s. with an expected P value of 0.50; mean correlation: $r = -0.038 \pm 0.030$, n.s.; Supplementary Fig. 13). This proportion was significantly lower than for cells in layer II ($Z = 9.5$, $P < 0.001$) or CA3 and CA1 ($Z = 5.5$ and $Z = 7.8$, respectively, both $P < 0.001$). The average phases at the beginning and end of grid fields in layer III were $195 \pm 81°$ and $192 \pm 123°$, respectively (circular means ± angular deviation; initial phase for layer II versus layer III: $F_{1,213} = 5.0$, $P < 0.05$, Watson-Williams test for two samples of angles). The mean slope of the linear regression line was $-0.078 \pm 0.283°\,\mathrm{cm}^{-1}$ (layer III versus layer II: $Z = 5.9$, $P < 0.001$). Taken together, these observations suggest that only a minority of the layer III cells exhibit phase precession.

To determine whether entorhinal phase precession is inherited from phase-precessing cells in the hippocampus, we measured the firing phases of grid cells after inactivation of the hippocampus by local infusion of the $GABA_A$ receptor agonist 5-aminomethyl-3-hydroxyisoxazole (muscimol) in four animals (Fig. 4; 132 fields before inactivation, 68 fields after inactivation). Place cells in dorsal and intermediate parts of the hippocampus were completely silenced by the infusion (Supplementary Fig. 14). On the test day, rats with recording electrodes in both dorsal hippocampus and entorhinal cortex first ran for 10 min in the drug-free state. Grid cells in layer II expressed strong phase precession, as in the main experiment (Fig. 4a, c; phase was negatively correlated with position in 107 of 132 fields (81%)). Muscimol was then infused; 5–15 min later, the rat was placed back and recording was resumed. Phase precession was assessed from the moment when the mean firing rate of the hippocampal place cells passed below 1% of the baseline activity (11–30 min after the start of the infusion). Hippocampal inactivation did not disrupt the spatially confined firing of the grid cells during the recording period, although the fields became wider and less stable (Fig. 4b and Supplementary Fig. 14). Theta oscillations and phase precession were still present (Fig. 4b, d). There was no decrease in the number of fields with a negative phase versus position correlation (before: 81%; after: 82%, $Z = 0.57$, n.s.) and the mean correlation between position and phase was not significantly reduced (before:

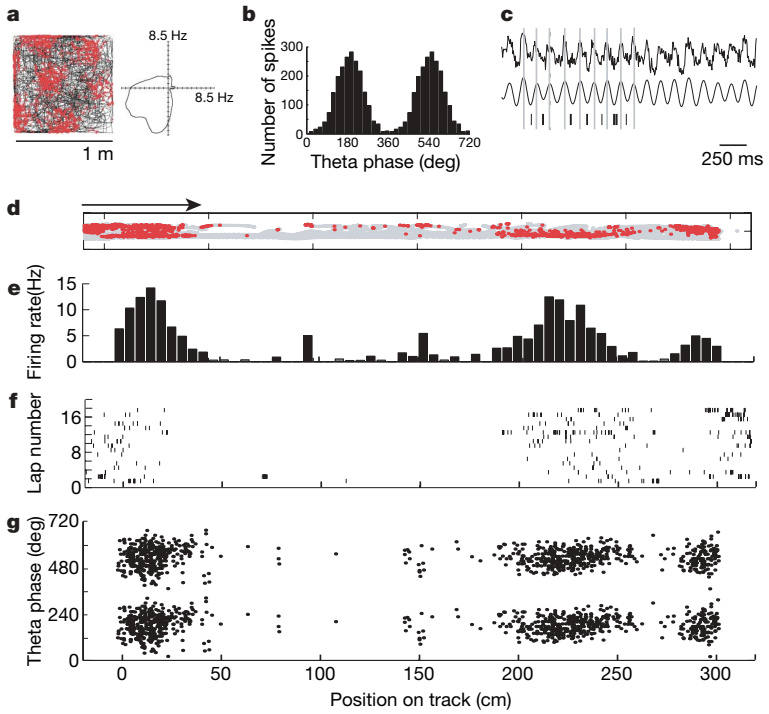


Figure 2 | Phase locking in layer III of MEC. **a**, Firing field (left) and directional tuning (right) of a layer III grid cell in the open field. **b**, Phase distribution for the cell in **a**. **c**, Entorhinal EEG with spike times for the same layer III cell during 2.3 s of track running. **d**–**g**, Trajectory with spike positions (**d**), linearized rate maps (**e**), raster plot for successive laps (**f**), and theta phase as a function of position (**g**). All symbols and procedures are as in Fig. 1. Note phase locking to the trough of the theta cycle. For additional examples see Supplementary Fig. 3.

$r = -0.27 \pm 0.03$; after: $r = -0.20 \pm 0.03$, $t_{198} = 1.6$, $P > 0.10$). Taken together, these observations suggest that phase precession in grid cells originates independently of the hippocampus.

We finally examined whether entorhinal phase precession is accompanied and caused by experience-dependent changes in the shape and the size of the firing fields, as proposed for hippocampal place cells[17,18]. Reflecting a possible development of asymmetric synaptic potentiation during repeated running along a fixed trajectory, expansion of the firing field in the direction opposite to the animal's movement may be part of a hebbian sequence coding process[19,20]. During running on the track, grid fields generally shifted against the running direction (Supplementary Figs 15–17). Most grid fields were skewed with the longer tail at the entrance of the field, but the asymmetry was expressed from the first lap, with no further development during the trial (Supplementary Fig. 15). Phase precession was not dependent on asymmetrization or centre-of-mass displacement (phase–position correlation versus centre-of-mass shift: $r = -0.02$; phase–position correlation versus skewness: $r = 0.06$; phase–position correlation versus field width: $r = -0.06$; 133 fields in layer II, all $P > 0.50$; Fig. 1f, g, and Supplementary Fig. 18). The apparent independence of phase precession and field asymmetry is consistent with the persistence of phase precession under conditions that block asymmetrization of place fields in the hippocampus[21].

The key findings of this study are twofold: first, the presence of strong theta-phase precession in one of the major excitatory inputs to place cells in the hippocampus, the projection from layer II of MEC, and second, the fact that phase precession in this region originates independently of the hippocampus. Phase precession in grid cells persisted after hippocampal inactivation, suggesting that the effect is not inherited from place cells. Instead, the phase advance is likely to originate locally, possibly in the principal cells of layer II, which express the strongest progression. Phase precession in layer II cells may be passed on to occasional grid cells in layer III as well as to place cells in CA3, which might convey the effect further to place cells in CA1. The fact that hippocampal phase precession persists when the local theta rhythm is reset by strong intrahippocampal inhibition[22] is consistent with an extrahippocampal source.

The observation of phase precession in grid cells has implications for the underlying neural mechanism. The data confirm some predictions of the dual-oscillator model for grid formation, in which spatial oscillations are proposed to result from interference between the global theta oscillation and speed-dependent direction-modulated subcellular oscillations with a slightly higher theta frequency[12,13]. Phase precession is a necessary consequence of interference between such signals. The differential expression of phase precession across entorhinal cell layers matches the variation in voltage-dependent subthreshold membrane potential oscillations in that stellate cells in layer II[23–25] and pyramidal cells in layer V[26,27] resonate at theta frequency, whereas layer III pyramidal cells seem not to display such intrinsic rhythmicity, at least not in brain slices[25,28]. Together with the link between grid size and intrinsic theta frequency in layer II cells[29], this suggests that phase precession and grid formation may be mechanistically associated with subthreshold membrane potential oscillations and wave interference. However, the match does not rule out other layer-specific intracellular or neural-network mechanisms, such as the interaction of an external input at theta frequency with a slow monotonically increasing excitatory conductance[3,4]. The nonlinear nature of the precession and the asymmetry of the fields do not follow from the dual-oscillator model, whereas asymmetric inputs could potentially explain these properties[4], if the plasticity is allowed to take place during neural development[18]. The bimodality of the phase–position relationship suggests that a combination of mechanisms may be involved[16], although at present only wave interference accounts simultaneously for phase precession and spatial periodicity.

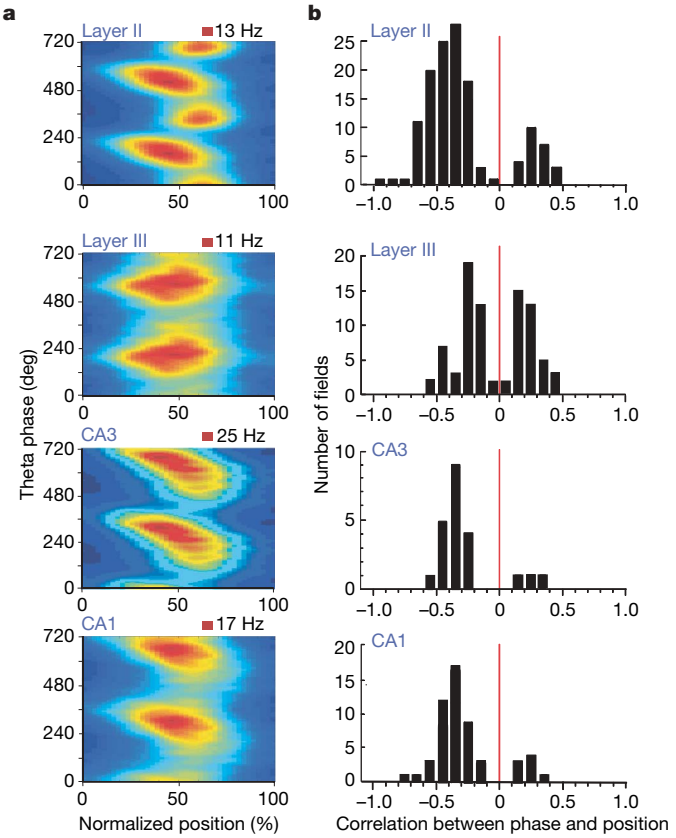


Figure 3 | Population data. a, Colour-coded spatiotemporal firing fields for layers II and III of MEC and CA3 and CA1 of the hippocampus. Each map shows the mean firing rate (colour) as a function of position (x) and phase (φ)[4]. Individual fields were rescaled to 100 pixels. Phase was defined relative to the local theta oscillation (not normalized). Red is maximum; blue is minimum. Peak rates are indicated. Note the strong phase advance in layer II, CA3 and CA1 but not in layer III. Also note the bimodal position–phase relationship in layer II (for individual examples see Supplementary Fig. 2). **b**, Frequency histograms showing the distribution of phase precession (same groups as in **a**). The strength of phase precession was expressed for each firing field by the correlation between phase and position at the phase rotation that gave the regression line with the largest explained variance. The distributions were bimodal with 0 as a local minimum because the selection criterion favoured rotations with high explained variance (and high correlations).

METHODS SUMMARY

Neuronal activity was recorded from grid cells and place cells in layer II, layer III or layer V of the dorsocaudal MEC and/or CA3 or CA1 of the dorsal hippocampus[10,11] in 23 male Long–Evans rats. Spike-triggered activity and local EEG were sampled in blocks of 10 min while the rat ran back and forth on a 235-cm or 320-cm linear track with food available at the turning points. Four animals were tested after inactivation of the hippocampus by bilateral infusion of the γ-aminobutyric acid$_A$ (GABA$_A$) receptor agonist muscimol into the region of dorsal hippocampus that provides the strongest output to the recording area in MEC[30]. Every spike was assigned a phase relative to the local theta oscillation, with 0° and 180° referring to the peak and the trough, respectively, of the theta wave. Phase change was expressed for each population of firing fields by plotting the mean firing rate as a function of position (x) and phase (φ) (Fig. 3a)[4]. Phase precession was assessed by parametrically rotating the phase by position (or time) matrix in steps of 1° across the phase cycle (360°) and fitting a linear regression line for each rotation[1,4] (Supplementary Fig. 10). The correlation for the rotation that gave the regression line with the largest explained variance R^2 (regression sum of squares divided by total sum of squares) was taken as an indicator of the strength of the phase precession. The slope of the best-fit regression line was used to estimate the rate of phase precession.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

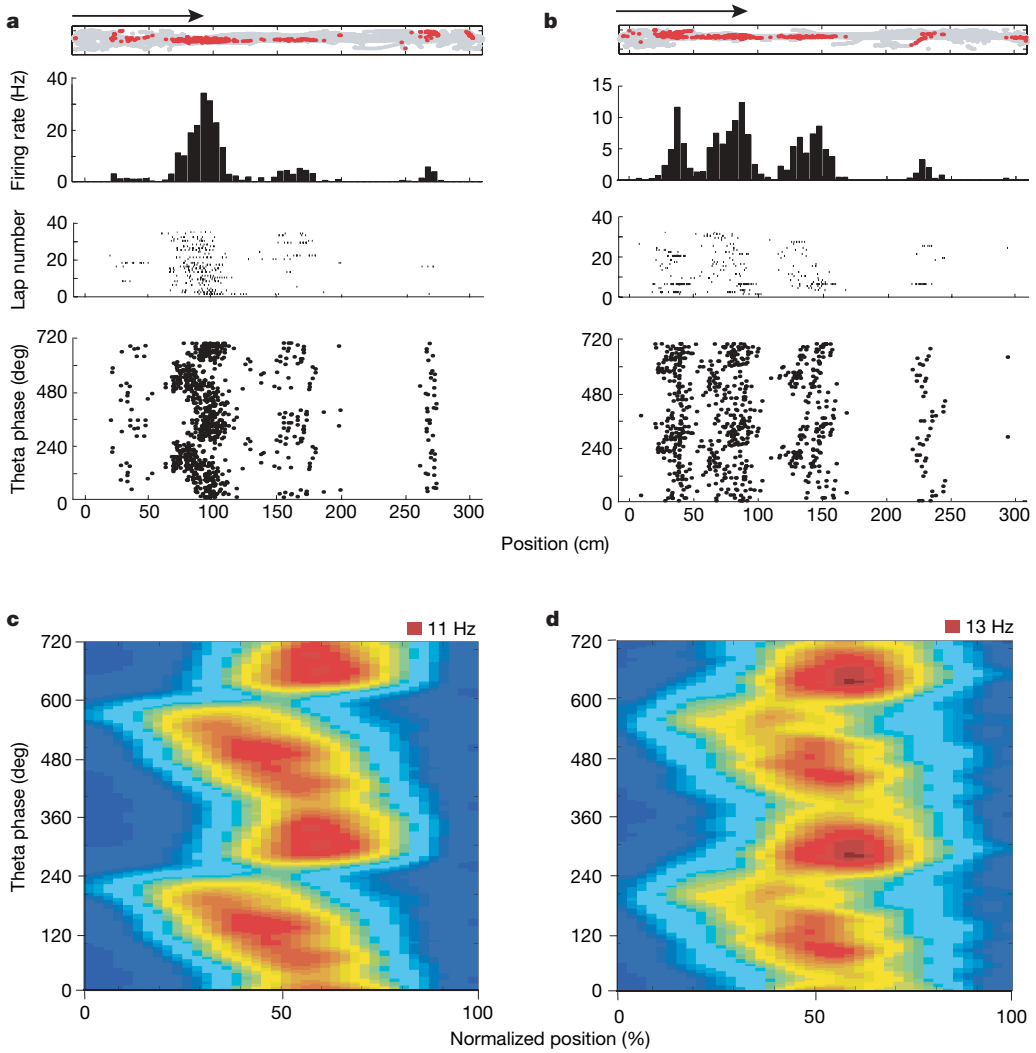


Figure 4 | Phase precession in entorhinal cortex after inactivation of the hippocampus. **a**, **b**, Firing fields and theta phase of an entorhinal grid cell recorded simultaneously with hippocampal place cells on the linear track before (**a**) and after (**b**) inactivation of the hippocampus. All place cells were silent at the time when phase precession was estimated (Supplementary Fig. 14). Top to bottom: trajectory with spike positions, linearized rate maps, raster plots for successive laps, and theta phase as a function of position. **c**, **d**, Spatiotemporal firing fields for the entire sample of grid fields from layer II cells recorded before (**c**) and after (**d**) inactivation of the hippocampus. Symbols are as in Figs 1 and 3. Note the persistence of phase precession after hippocampal inactivation.

Received 28 December 2007; accepted 1 April 2008.
Published online 14 May 2008.

1. O'Keefe, J. & Recce, M. L. Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus* **3,** 317–330 (1993).
2. Skaggs, W. E., McNaughton, B. L., Wilson, M. A. & Barnes, C. A. Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus* **6,** 149–172 (1996).
3. Harris, K. D. *et al.* Spike train dynamics predicts theta-related phase precession in hippocampal pyramidal cells. *Nature* **417,** 738–741 (2002).
4. Mehta, M. R., Lee, A. K. & Wilson, M. A. Role of experience and oscillations in transforming a rate code into a temporal code. *Nature* **417,** 741–746 (2002).
5. Huxter, J., Burgess, N. & O'Keefe, J. Independent rate and temporal coding in hippocampal pyramidal cells. *Nature* **425,** 828–832 (2003).
6. Tsodyks, M. V., Skaggs, W. E., Sejnowski, T. J. & McNaughton, B. L. Population dynamics and theta rhythm phase precession of hippocampal place cell firing: a spiking neuron model. *Hippocampus* **6,** 271–280 (1996).
7. Jensen, O. & Lisman, J. E. Hippocampal CA3 region predicts memory sequences: accounting for the phase precession of place cells. *Learn. Mem.* **3,** 279–287 (1996).
8. Dragoi, G. & Buzsáki, G. Temporal encoding of place sequences by hippocampal cell assemblies. *Neuron* **50,** 145–157 (2006).
9. Jones, M. W. & Wilson, M. A. Phase precession of medial prefrontal cortical activity relative to the hippocampal theta rhythm. *Hippocampus* **15,** 867–873 (2005).
10. Fyhn, M., Molden, S., Witter, M. P., Moser, E. I. & Moser, M.-B. Spatial representation in the entorhinal cortex. *Science* **305,** 1258–1264 (2004).
11. Hafting, T., Fyhn, M., Molden, S., Moser, M.-B. & Moser, E. I. Microstructure of a spatial map in the entorhinal cortex. *Nature* **436,** 801–806 (2005).
12. O'Keefe, J. & Burgess, N. Dual phase and rate coding in hippocampal place cells: theoretical significance and relationship to entorhinal grid cells. *Hippocampus* **15,** 853–866 (2005).
13. Burgess, N., Barry, C. & O'Keefe, J. An oscillatory interference model of grid cell firing. *Hippocampus* **17,** 801–812 (2007).
14. Blair, H. T., Welday, A. C. & Zhang, K. Scale-invariant memory representations emerge from moire interference between grid fields that produce theta oscillations: a computational model. *J. Neurosci.* **27,** 3211–3229 (2007).
15. Sargolini, F. *et al.* Conjunctive representation of position, direction and velocity in entorhinal cortex. *Science* **312,** 754–758 (2006).
16. Yamaguchi, Y., Aota, Y., McNaughton, B. L. & Lipa, P. Bimodality of theta phase precession in hippocampal place cells in freely running rats. *J. Neurophysiol.* **87,** 2629–2642 (2002).
17. Mehta, M. R., Barnes, C. A. & McNaughton, B. L. Experience-dependent, asymmetric expansion of hippocampal place fields. *Proc. Natl Acad. Sci. USA* **94,** 8918–8921 (1997).
18. Mehta, M. R., Quirk, M. C. & Wilson, M. A. Experience-dependent asymmetric shape of hippocampal receptive fields. *Neuron* **25,** 707–715 (2000).
19. Blum, K. I. & Abbott, L. F. A model of spatial map formation in the hippocampus of the rat. *Neural Computat.* **8,** 85–93 (1996).

20. Mehta, M. R. Neuronal dynamics of predictive coding. *Neuroscientist* **7,** 490–495 (2001).
21. Ekstrom, A. D., Meltzer, J., McNaughton, B. L. & Barnes, C. A. NMDA receptor antagonism blocks experience-dependent expansion of hippocampal 'place fields'. *Neuron* **31,** 631–638 (2001).
22. Zugaro, M. B., Monconduit, L. & Buzsaki, G. Spike phase precession persists after transient intrahippocampal perturbation. *Nature Neurosci.* **8,** 67–71 (2005).
23. Alonso, A. & Llinas, R. R. Subthreshold Na^+-dependent theta-like rhythmicity in stellate cells of entorhinal cortex layer II. *Nature* **342,** 175–177 (1989).
24. Klink, R. & Alonso, A. Ionic mechanisms for the subthreshold oscillations and differential electroresponsiveness of medial entorhinal cortex layer II neurons. *J. Neurophysiol.* **70,** 144–157 (1993).
25. Erchova, I., Kreck, G., Heinemann, U. & Herz, A. V. Dynamics of rat entorhinal cortex layer II and III cells: characteristics of membrane potential resonance at rest predict oscillation properties near threshold. *J. Physiol. (Lond.)* **560,** 89–110 (2004).
26. Schmitz, D., Gloveli, T., Behr, J., Dugladze, T. & Heinemann, U. Subthreshold membrane potential oscillations in neurons of deep layers of the entorhinal cortex. *Neuroscience* **85,** 999–1004 (1998).
27. Hamam, B. N., Kennedy, T. E., Alonso, A. & Amaral, D. G. Morphological and electrophysiological characteristics of layer V neurons of the rat medial entorhinal cortex. *J. Comp. Neurol.* **418,** 457–472 (2000).
28. Dickson, C. T., Mena, A. R. & Alonso, A. Electroresponsiveness of medial entorhinal cortex layer III neurons *in vitro*. *Neuroscience* **81,** 937–950 (1997).
29. Giocomo, L. M., Zilli, E. A., Fransen, E. & Hasselmo, M. E. Temporal frequency of subthreshold oscillations scales with entorhinal grid cell field spacing. *Science* **315,** 1719–1722 (2007).
30. Kloosterman, F., Witter, M. P. & Van Haeften, T. Topographical and laminar organization of subicular projections to the parahippocampal region of the rat. *J. Comp. Neurol.* **455,** 156–171 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Skjerpeng for programming; M. P. Witter for advice on electrode localization; A. Treves for statistical advice; A. M. Amundsgaard, D. Derdikman, K. Haugen, K. Jenssen, E. Sjulstad and H. Waade for technical or other assistance; and several colleagues for discussion. The work was supported by the Kavli Foundation, a Centre of Excellence grant from the Norwegian Research Council and the 2006 Life Science award of the Fondation Bettencourt Schueller.

Author Contributions T.H., M.F., M.-B.M. and E.I.M. planned the experiments; T.H., M.F., T.B. and M.-B.M. performed the experiments; all authors analysed the data; and E.I.M. wrote the paper. All authors discussed the results and contributed to the manuscript. M.F. and T.H. contributed equally.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to E.I.M. (edvard.moser@ntnu.no).

METHODS

Implants and data collection. Twenty-three rats were anaesthetized and chronically implanted with microdrives connected to four platinum-plated tetrodes of twisted 17-μm HM-L coated platinum-iridium wire[10,11]. The tetrodes were implanted above the dorsocaudal MEC, 4.5 mm lateral to the midline (ML) and 0.2–0.3 mm anterior to the transverse sinus, at an angle of about 10° in the anterior direction. Animals received one microdrive in MEC of each hemisphere, at corresponding positions, or one in MEC and one in CA3 (anteroposterior (AP), −3.1; ML, 2.8) or CA1 (AP, −4.0; ML, 3.1). Simultaneous recordings were obtained from left and right MEC in eight animals, from MEC and CA1 in seven animals, and from MEC and CA3 in four animals.

After recovery, the rats were given daily sessions of cell screening and pre-training in a square enclosure (100 cm × 100 cm; 50 cm high) and on a linear track (235 or 320 cm long, 10 cm wide). Single units and EEG were recorded by means of alternating-current-coupled unity-gain operational amplifiers on the rat's head, with a counterbalanced cable[10]. Position was recorded by tracking a red light-emitting diode on the headstage[10]. The tetrodes were lowered in steps of 50 μm until multiple well-separated theta-modulated large-amplitude low-frequency neurons appeared at depths of about 2.0 mm or more in the dorsolateral MEC or the hippocampus. The rat was then moved to the linear track for further recording. Trials were 10–30 min, each consisting of a minimum of 15 laps (Supplementary Methods). After each session on the track, the tetrodes were moved further until new well-separated grid cells were encountered in the box.

Analysis. Cells were separated with the use of conventional cluster-cutting methods (Supplementary Methods). Unsmoothed linearized rate maps were constructed by plotting firing rate along the length of the track, with a bin size of 5 cm. Firing fields were identified in two steps. First, a preliminary field was defined as any contiguous region of at least 15 cm (three bins) where the firing rate of the cell was above 10% of the peak rate. The final field was then defined by extending the preliminary field successively across bins from each end, starting from the 10% border, until a bin was reached in which either the rate was higher than in the preceding bin, or the rate was lower than 1% of the peak rate. Fields with less than 50 spikes, fields that included the turning points of the track, and fields with spatial coherence (spatial correlation between neighbouring bins) of less than 0.70 were discarded. Epochs with instantaneous running speeds of less than 10 cm s^{-1} were excluded. Outward and inward directions on the track were analysed separately.

The local EEG was filtered by an acausal (zero phase shift) fast Fourier transform bandpass filter, with 5–6 Hz and 10–11 Hz as cut-off frequencies (Supplementary Methods). Every recorded spike was assigned a phase relative to the filtered signal, with 0° and 180° referring to peak and trough, respectively. Spike times between peaks and troughs were interpolated linearly. The phases at the entrance and the exit of the firing field were defined as the mean phase of the field's first and last ten spikes, respectively, on the position axis, irrespective of lap number. The strength and rate of precession were determined by parametrically rotating the phase by position (or time) matrix for each cell[1,4]. Phase rows were shifted in steps of 1° across the entire phase cycle, from 0° to 360°, continuing across the 360° border for all rotations larger than 0°. For each rotation, a linear regression curve was fitted. The correlation between phase and position for the rotation that gave the regression line with the largest explained variance R^2 was taken as an indicator of the strength of phase precession (Supplementary Methods and Supplementary Fig. 10). The cell's rate of precession was estimated as the slope of the regression line at this rotation. Finally, to estimate the overall phase change by an independent method not restrained by assuming linear and unimodal distributions, the overall phase change was expressed for each cell population by plotting the mean firing rate in the population as a function of position (x) and phase (φ) across 40 bins along the position axis and 100 bins along the phase axis[4]. Phase–position fields were stacked on top of each other and the mean rate was computed for each population. Smoothing was obtained by integrating over each phase–position bin a gaussian factor that spread the contribution of each spike to the adjacent bins, using bin widths of 1 for both x and φ. Firing rate was colour-coded, with blue as the minimum and red as the maximum.

Hippocampal inactivation. The effect of hippocampal output on phase precession in entorhinal cortex was tested in four rats by inactivating the hippocampus by local infusion of the $GABA_A$ receptor agonist muscimol (Sigma) through 26-gauge outer cannulae (C315G; Plastics One) above each dorsal hippocampus (AP 2.5 mm, ML ± 2.5 mm, dorsoventral 1.6 mm at an angle of 30° in the posterior direction). Microdrives were positioned in hippocampus (AP 4.5, ML 3.6) and MEC (as in the main experiment). Between three and five weeks after the start of training on the track, place cells and grid cells were first recorded for 10 min in the absence of drug. Muscimol dissolved in PBS (pH 7.4; 0.5 μg μl^{-1}) was then infused bilaterally at 0.08 μl min^{-1} by means of two 33-gauge internal cannulae (C315I; Plastics One) connected by polyethylene tubing to a 25-μl syringe in a CMA/100 infusion pump. The tip of the inner cannula protruded 0.9 mm out of the implanted outer cannula. The total volume of each hippocampal infusion was 0.25 μl (five experiments, three rats) or 0.64 μl (three experiments, two rats). The internal cannulae were retracted 2 min after each infusion. The rat was then placed back on the track and recording was resumed for 30 min (Supplementary Methods). The distribution of muscimol was tested in two animals without entorhinal recordings, one with electrodes at a more posterior pair of locations (AP 5.7, ML 4.7), about 50–60% along the dorsoventral axis of the hippocampus (Supplementary Fig. 14).

LETTERS

Rapid strengthening of thalamo-amygdala synapses mediates cue–reward learning

Kay M. Tye[1,2], Garret D. Stuber[1], Bram de Ridder[1], Antonello Bonci[1,2,3,4] & Patricia H. Janak[1,2,3,4]

What neural changes underlie individual differences in goal-directed learning? The lateral amygdala (LA) is important for assigning emotional and motivational significance to discrete environmental cues[1–4], including those that signal rewarding events[5–8]. Recognizing that a cue predicts a reward enhances an animal's ability to acquire that reward; however, the cellular and synaptic mechanisms that underlie cue–reward learning are unclear. Here we show that marked changes in both cue-induced neuronal firing and input-specific synaptic strength occur with the successful acquisition of a cue–reward association within a single training session. We performed both *in vivo* and *ex vivo* electrophysiological recordings in the LA of rats trained to self-administer sucrose. We observed that reward-learning success increased in proportion to the number of amygdala neurons that responded phasically to a reward-predictive cue. Furthermore, cue–reward learning induced an AMPA (α-amino-3-hydroxy-5-methyl-isoxazole propionic acid)-receptor-mediated increase in the strength of thalamic, but not cortical, synapses in the LA that was apparent immediately after the first training session. The level of learning attained by individual subjects was highly correlated with the degree of synaptic strength enhancement. Importantly, intra-LA NMDA (*N*-methyl-D-aspartate)-receptor blockade impaired reward-learning performance and attenuated the associated increase in synaptic strength. These findings provide evidence of a connection between LA synaptic plasticity and cue–reward learning, potentially representing a key mechanism underlying goal-directed behaviour.

Basolateral amygdala (BLA) neurons are phasically responsive to reward-predictive cues[8–11], which is consistent with the idea that cue-evoked neuronal firing emerges as a consequence of cue–reward associations. The BLA is composed of multiple nuclei, including the LA, the first site of convergence for sensory inputs carrying information about conditioned and unconditioned stimuli to the amygdala[1,4,12–15]. Thus, the LA is a likely initial site for the formation of cue–reward associations that endow the cue with motivational significance that impacts on reward-seeking behaviour.

To test the hypothesis that successful acquisition of cue-directed reward-seeking behaviour is dependent on neuronal plasticity in the LA, we examined LA neuronal firing in response to a reward-predictive cue during training on a sucrose self-administration task (Supplementary Fig. 1). To control for neural activity associated with the motor output of operant responding, and to ensure that the sensory cue predicted reward delivery and not the operant response alone, beam breaks at a nose-poke response port ('nose-poke operandum') were reinforced with a cue and sucrose reward after about 50% of nose-pokes (Fig. 1a, b). In rats that successfully acquired this task (see Methods), about half of recorded neurons (49%; 60 of 122 neurons from seven rats during the first session in which each rat met the acquisition criterion) that did not respond to the cue before acquisition developed a robust phasic response to cue onset with acquisition (Fig. 1c, d, and Supplementary Fig. 2). Cue encoding increased across sessions: the cue-evoked population response of all neurons recorded in the third session was enhanced relative to the first session (session × time interaction, $F_{9,1944} = 4.15$, $P < 0.0001$), specifically within the 50 ms after cue onset ($P < 0.003$; Fig. 1e, Supplementary Fig. 3 and Supplementary Table 1). These changes over sessions were predictive of behaviour: increasing proportions of neurons were recruited to encode the reward-predictive cue as individual rats improved reward-learning performance (Fig. 1f, g). Task efficiency, a behavioural index defined as the number of rewards earned divided by the number of cues presented, and task accuracy, a behavioural index defined as the difference in the number of correct and incorrect port entries divided by the total number of port entries, were significantly correlated ($P < 0.0001$ and $P = 0.0066$, respectively; Supplementary Table 2) with the percentage of neurons per rat that showed phasic responses to the reward-predictive cue (Fig. 1f, g). Control studies confirmed that the increase in cue encoding is specific to acquisition of the cue–reward association and is not due to non-associative factors, such as sensitization (Supplementary Figs 4 and 5). These data demonstrate that development of cue-evoked responses in the LA depends on the acquired reward-predictive nature of the cue. Further, the greater the proportion of neurons recruited to encode the reward-predictive cue, the better the rat learned the cue–reward association, and the more successful the rat was at earning rewards.

Because our *in vivo* recordings showed rapidly occurring changes in cue-related firing in the LA during successful cue–reward learning, we proposed that the mechanism underlying these changes was an increase in synaptic strength of thalamic or cortical sensory afferents onto LA neurons; we tested this hypothesis with *ex vivo* experimentation (Supplementary Fig. 6). Rats were trained on a single session of the same behavioural model and classified as learners (top 50%) or non-learners (bottom 50%) as defined by our learning indices of task efficiency and task accuracy (Supplementary Fig. 7). Any unearned sucrose was delivered in the home cage immediately after the session, ensuring that all rats received the same amount of sucrose. Brains were collected about 30 min after the end of the session for the preparation of acute slices of the LA. We stimulated the internal or external capsule to evoke excitatory postsynaptic currents (EPSCs) from thalamic or cortical afferents, respectively, and used whole-cell patch-clamp techniques within visually identified pyramidal neurons to measure EPSCs containing AMPA receptor (AMPAR)-mediated and NMDA receptor (NMDAR)-mediated currents. We found that the AMPAR/NMDAR ratio, an index of glutamatergic synaptic strength[16,17], varied with task performance and afferent (main effects of group, $F_{2,29} = 11.01$, $P < 0.001$; afferent, $F_{1,29} = 22.13$, $P < 0.001$;

[1]Ernest Gallo Clinic and Research Center, University of California, San Francisco, Emeryville, California 94608, USA. [2]Program in Neuroscience, [3]Department of Neurology, and [4]Wheeler Center for the Neurobiology of Addiction, University of California, San Francisco, California 94143, USA.

group × afferent interaction, $F_{2,29} = 7.38$, $P < 0.004$) such that learners had a larger AMPAR/NMDAR ratio at thalamic synapses ($P < 0.001$; learners, 1.03 ± 0.04; non-learners, 0.58 ± 0.08; naives, 0.47 ± 0.05 (means ± s.e.m.)) but not cortical synapses (learners, 0.45 ± 0.08; non-learners, 0.46 ± 0.10; naives, 0.47 ± 0.04) in the LA relative to non-learners and naives, which did not differ from each other ($P = 0.84$; Fig. 2a, b). We determined the correlation between each rat's behavioural performance, as measured by either task efficiency or task accuracy, and the AMPAR/NMDAR ratio, and found a significant positive relationship at thalamic inputs ($P = 0.0003$ and $P = 0.006$, respectively) but not cortical inputs ($P = 0.89$ and $P = 0.55$, respectively; Fig. 2c–f and Supplementary Table 3). Hence, thalamo-amygdalar synaptic strength predicted the success of individual rats' reward-learning performance.

A change in the relative contribution of AMPARs and NMDARs to compound EPSCs may reflect an increase in AMPAR currents and/or a decrease in NMDAR currents at thalamo-amygdalar synapses. To determine whether AMPAR currents were modified during reward learning, we examined AMPAR-mediated miniature EPSCs (mEPSCs), which reflect spontaneously released vesicles of glutamate[18]. Typically, an increase in mEPSC amplitude indicates an increase in postsynaptic AMPAR number or function, whereas an increase in mEPSC frequency indicates an increase in the probability of transmitter release (P_r) or in the number of synapses[18]. mEPSC amplitude was related to task performance ($F_{2,29} = 30.75$, $P < 0.001$), with a greater mean amplitude from LA neurons of learners ($P < 0.001$; 15.88 ± 0.89 pA) than from those of non-learners (9.98 ± 0.29 pA) or naives (10.05 ± 0.39 pA), which did not differ from each other ($P = 0.87$; Fig. 3a, b, d). In contrast, the mean mEPSC frequency was not different ($F_{2,29} = 0.5$, $P = 0.61$) in learners (6.45 ± 1.48 Hz), non-learners (5.36 ± 1.16 Hz) and naives (4.96 ± 1.14 Hz) (Fig. 3a, c, e). To examine further whether learning altered P_r, we examined the paired-pulse ratio[19] (inter-stimulus interval 50 ms; Fig. 3f). There was no change in the paired-pulse ratio for

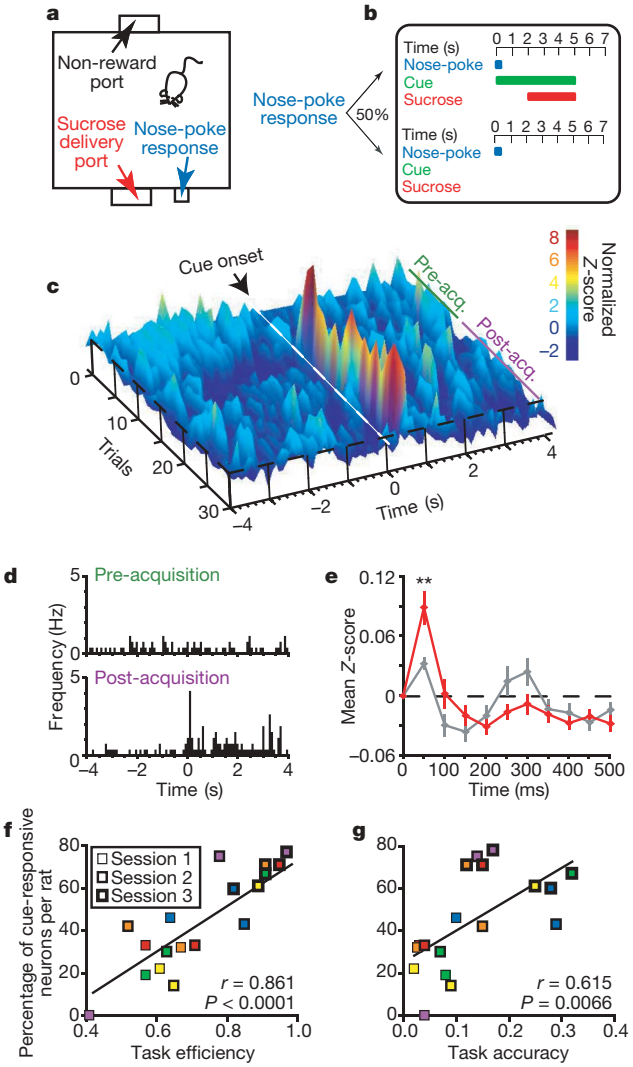


Figure 1 | Reward-related learning success is correlated with rapid increases in cue-related firing. **a**, Diagram of operant chamber from above. **b**, Schematic of behavioural paradigm. **c**, Temporal dynamics of neuronal population response to the reward-predictive cue. Spike activity of all simultaneously recorded LA units ($n = 13$) from a rat that successfully acquired the task during the first session; 100-ms bins. This population of neurons develops a response to the onset of the reward-predictive cue with task acquisition (acq.). **d**, Peri-event histogram of a single LA neuron from a different rat that successfully acquired the task in the first session; 29 trials in each epoch. **e**, Population histograms (50-ms bins, error bars indicate s.e.m.) of the mean *Z*-score for all neurons recorded during session 1 (grey; $n = 95$ neurons) and session 3 (red; $n = 123$ neurons) for all rats ($n = 7$). Two asterisks, $P < 0.004$. **f**, **g**, Correlation between proportion of cue-responsive neurons and (**f**) task efficiency and (**g**) task accuracy across three sessions. Colours indicate the same rat on different sessions. Only rats with at least six neurons per session were included in scatter plots ($n = 6$). For all peri-event histograms, time zero indicates cue onset.

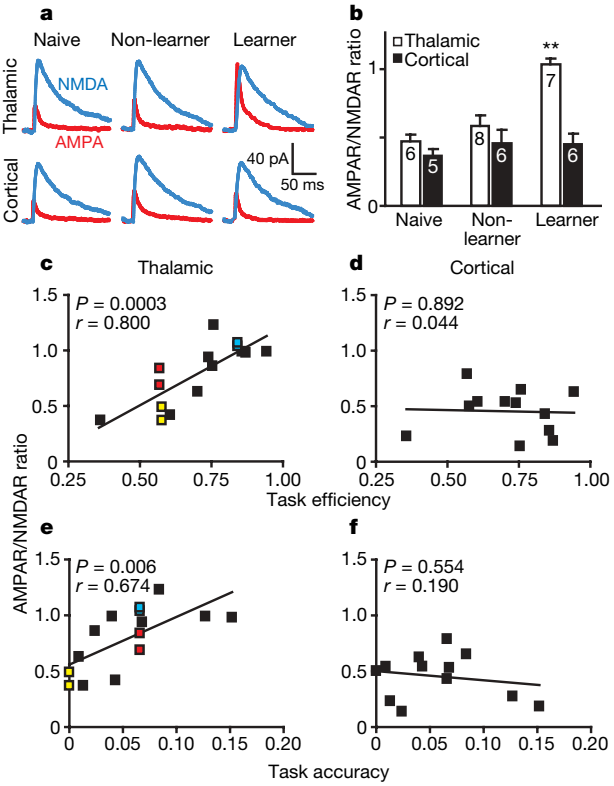


Figure 2 | Degree of AMPAR/NMDAR enhancement predicts cue–reward learning. **a**, EPSCs evoked by stimulation of thalamic or cortical afferents in rats that were naive, non-learners or learners. **b**, AMPAR/NMDAR ratios evoked from thalamic afferents were significantly increased in learners ($n = 6$ rats) in comparison with non-learners ($n = 6$ rats) or naives ($n = 5$ rats). Numbers in bars indicate numbers of cells; error bars indicate s.e.m. Two asterisks, $P < 0.001$, significant difference from other groups as well as from cortical afferent. **c**–**f**, Correlation between AMPAR/NMDAR ratio and either task efficiency (**c**, **d**) or task accuracy (**e**, **f**) for EPSCs evoked from thalamic (**c**, **e**), but not cortical (**d**, **f**), pathways; the subjects were the same as in **b**. Colours indicate multiple cells recorded from the same rat; black indicates single cells recorded from each rat.

either afferent ($F_{1,33} = 0.02$, $P = 0.89$) among naives, non-learners or learners (main effect of group, $F_{2,33} = 0.35$, $P = 0.71$; group × afferent interaction, $F_{2,33} = 0.40$, $P = 0.67$), indicating that learning does not cause an immediate change in P_r and that the rapid increase in AMPAR/NMDAR ratio is mediated postsynaptically.

The induction of associative long-term potentiation in the LA depends on the activation of NMDARs[20,21], which can lead to increases in AMPAR currents[18]. In addition, NMDAR blockade within the BLA impairs acquisition, but not performance, in two similar appetitive tasks[22,23]. To test whether the learning-induced synaptic changes we observed are dependent on NMDAR activation, we locally infused the NMDAR antagonist AP5 (3 μg per side) or vehicle (artificial cerebrospinal fluid; aCSF) into the LA bilaterally before training (Supplementary Fig. 8). To control for the possibility that synaptic changes might be secondary to, rather than causal for, reduced behavioural performance, we included a third group in which rats received unilateral intra-LA infusions of AP5 and contralateral infusions of aCSF to provide a within-animal control. Task efficiency was impaired by AP5 ($F_{2,12} = 9.03$, $P < 0.005$) after both bilateral ($P < 0.007$) and unilateral ($P < 0.018$) intra-LA pre-training

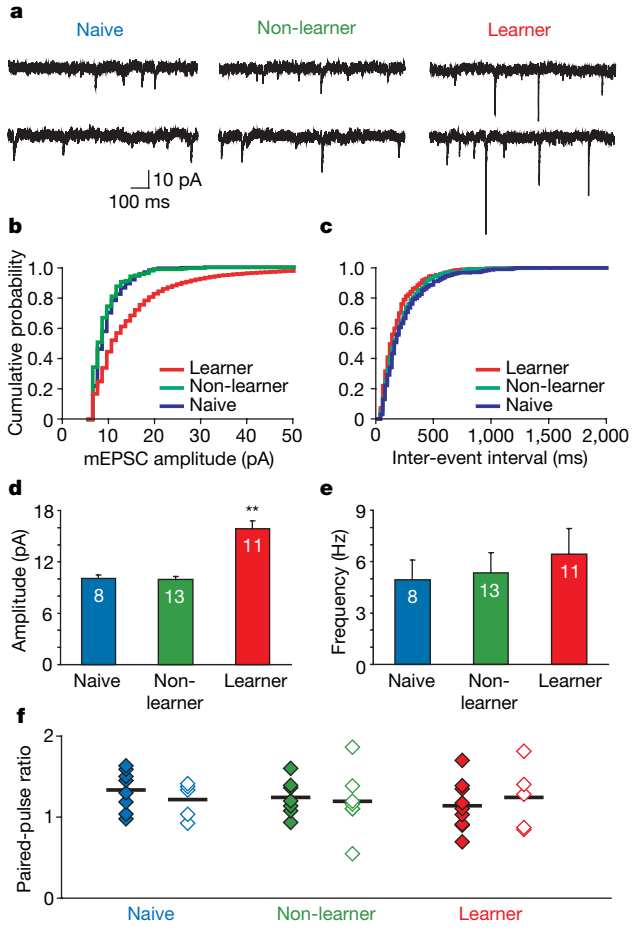


Figure 3 | Successful cue–reward learning induces an increase in mEPSC amplitude but not in frequency or paired-pulse ratio. **a**, Sample mEPSCs from rats that were naive, non-learners or learners. **b**, **c**, Cumulative probability plots of amplitude (**b**) and frequency (**c**) for representative neurons from each group; bins were 1 pA (**b**) and 20 ms (**c**). **d**, **e**, Learners ($n = 6$ rats) had increased mEPSC amplitude (**d**), but not frequency (**e**), relative to non-learners ($n = 6$ rats) and naive rats ($n = 5$ rats). Numbers in bars indicate numbers of cells; errors bars indicate s.e.m. Two asterisks, $P < 0.001$. **f**, Lack of change in paired-pulse ratio between the different groups of rats. Filled diamonds, ratios evoked from the thalamic pathway; open diamonds, ratios evoked from the cortical pathway; horizontal lines indicate the means.

infusions (Fig. 4a, c); bilateral, but not unilateral, intra-LA infusions of AP5 also impaired task accuracy ($F_{2,12} = 7.38$, $P < 0.009$; aCSF versus bilateral AP5, $P < 0.009$; Fig. 4b, c). The effect of AP5 was not attributable to a spread of drug into the neighbouring central nucleus of the amygdala (Supplementary Fig. 9).

After intra-LA infusions and the training session, brains from these rats were collected for the preparation of acute slices. Rats that received bilateral intra-LA infusions of AP5 showed a lower mean amplitude of mEPSCs ($P = 0.003$; 10.26 ± 0.41 pA; Fig. 4d, h) than after infusions with aCSF (13.09 ± 0.68 pA; Fig. 4e, h), whereas there was no change in mEPSC frequency between groups ($P = 0.66$; Fig. 4d, e, i). The decrease in task efficiency and the decrease in mEPSC amplitude after local infusion of an NMDAR antagonist suggest that cue–reward learning and the associated increase in AMPAR number or function are dependent on NMDAR activation. By comparing mEPSCs from rats with unilateral intra-LA AP5 infusions and contralateral aCSF infusions, we were able to determine with confidence that any differences between LA neurons treated with AP5 or aCSF are due to local NMDAR blockade rather than to an AP5-induced difference in task performance. Within subjects, we found that the amplitude of LA mEPSCs recorded after AP5 infusion into the LA on one side was significantly lower ($P < 0.001$; Fig. 4f, j) relative to aCSF infusion on the contralateral side (Fig. 4g, j), whereas there was no difference in frequency ($P = 0.99$; Fig. 4f, g, k). Local NMDAR blockade therefore attenuates the learning-dependent increase in postsynaptic AMPAR currents and impairs the acquisition of reward-directed behaviour.

These results show that, with cue–reward learning, cue-responsive neurons are rapidly recruited *in vivo*, thalamo-amygdalar synapses are selectively strengthened, and LA neurons show NMDAR-dependent increases and associated potentiation of AMPAR number or function. The proportion of cells recorded *in vivo* that developed a response to the reward-predictive cue is less than the proportion of cells that showed enhanced synaptic strength with learning (Fig. 2c). This suggests that the integration of multiple inhibitory and excitatory synapses on a given cell may constrain cue-related spike firing[3,24,25], even if that cell possesses enhanced thalamic inputs. The thalamic pathway is under strong inhibitory suppression[20,24] *in vivo*, whereas our *ex vivo* recordings were performed under γ-aminobutyric acid $(GABA)_A$-receptor antagonism to isolate EPSCs.

The parallel emergence of increased synaptic strength and cue-related firing in the LA neurons during reward learning suggests that this excitatory synaptic increase contributes to enhanced spike activity of LA neurons in response to the conditioned stimulus, driven by auditory and visual thalamic inputs that terminate in the LA[1,12]. Consistent with our results, auditory fear conditioning, which requires an intact LA[1,2,4], increases neuronal firing in response to a shock-predictive cue and potentiates transmission at thalamo-amygdalar synapses[26] by an NMDAR-dependent mechanism, probably a result of postsynaptic AMPAR trafficking[27]. Previous work on fear conditioning suggests that plasticity also occurs at cortical[28] synapses in the LA, although this enhancement was found at later time points than tested here. Single-unit recordings in the LA show that the thalamic pathway conditions more rapidly than the cortical pathway during fear conditioning[1,29]. Our findings, viewed in the context of fear conditioning, prompt further experimentation to determine whether rapidly occurring reward-learning-induced plasticity at thalamo-amygdalar synapses facilitates subsequent consolidation at other sites[30].

These findings indicate that rapid synaptic changes in the LA occur during the early stages of cue–reward learning. It is likely that this plasticity permits amygdala neurons to respond selectively to meaningful environmental stimuli and transmit this information to downstream brain regions for the expedited selection of an adaptive behavioural output.

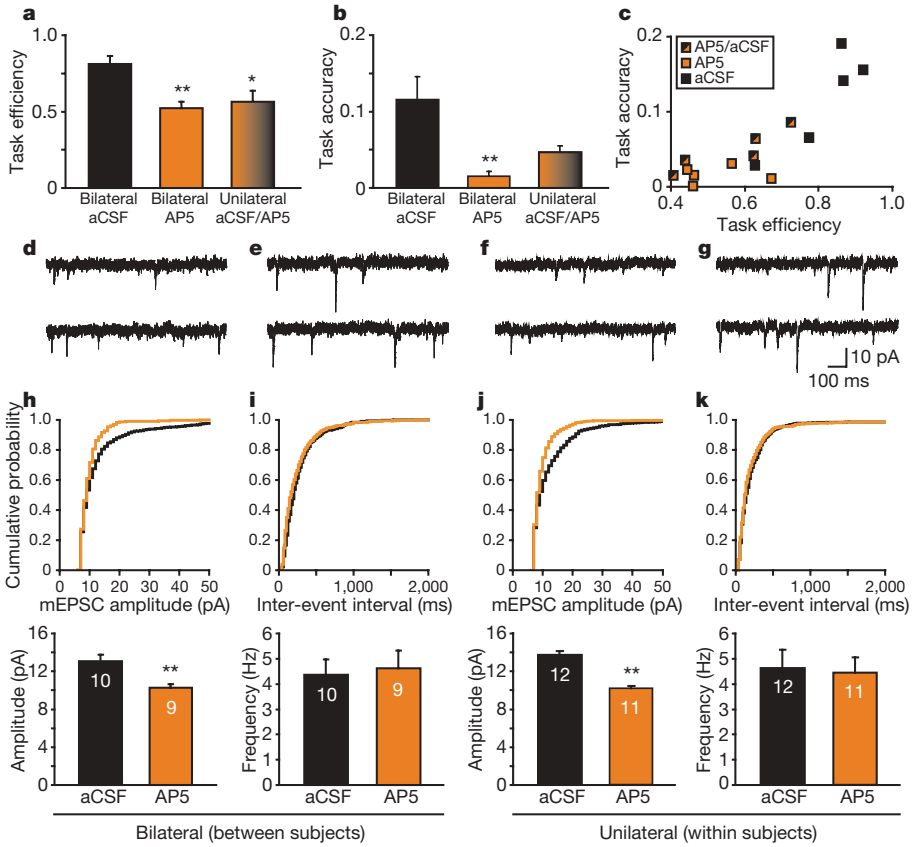


Figure 4 | Local NMDAR blockade attenuates reward-related learning and the associated increase in mEPSC amplitude. **a**, **b**, Measures of task performance among groups ($n = 5$ rats per group). Task efficiency (**a**) is decreased after either unilateral or bilateral AP5 intra-LA infusion, whereas task accuracy (**b**) is decreased after bilateral AP5 (asterisk, $P < 0.05$, two asterisks, $P < 0.009$, compared with aCSF). No other comparisons were significant. Error bars indicate s.e.m. **c**, Individual rat performances based on task efficiency and task accuracy. **d**–**g**, Sample mEPSCs from rats that received pre-training infusions. **d**, **e**, Traces recorded from rats that received bilateral infusions of AP5 (**d**) and aCSF (**e**). **f**, **g**, Traces recorded from a representative rat that received unilateral intra-LA infusions of AP5 (**f**) and aCSF (**g**). **h**–**k**, Cumulative probability plots of amplitude (**h**, **j**) and frequency (**i**, **k**) for mEPSCs in representative cells from rats receiving bilateral infusions of AP5 (orange) or aCSF (black) (**h**, **i**) or unilateral infusions of AP5 and aCSF (**j**, **k**). Below each probability plot is the corresponding bar graph indicating the group mean and s.e.m. There is a difference in amplitude (**h**, **j**), but not in frequency (**i**, **k**) for both bilaterally infused (**h**) and unilaterally infused (**j**) rats; two asterisks, $P < 0.001$, compared with aCSF. Numbers in bars indicate numbers of cells.

METHODS SUMMARY

Adult male Sprague–Dawley rats (250–350 g) were food-restricted to 90% of free-feeding body weight. Training session length was varied as follows: rats with chronic electrodes were trained daily for 3 h per session, for three sessions; rats with cannulae were trained for one 4-h session to allow the enhanced opportunity to express learning within a single session; and rats with no previous surgery were trained for one 2-h session, the median time required for task acquisition. Nose-poke responses were reinforced on a partial reinforcement schedule, with about 50% of responses reinforced by a 5-s compound light-tone stimulus, and 0.1 ml of 15% sucrose delivered 2 s after cue onset. Task acquisition was defined by more than 80% correct trials in a moving five-trial block. A correct trial was defined as a nose-poke yielding a cue presentation and subsequent port entry (within 10 s or before performing a different behaviour). Incorrect trials were defined as entering the port after a nose-poke without the cue. Phasic neuronal responses were deemed significant if the firing rate of a unit in any of five 100-ms bins in a 0–0.5-s response window after cue onset was different from the firing rate in a 0.5-s baseline epoch by a Wilcoxon signed-rank test. Group values are expressed as means ± s.e.m. The statistical significance of multiple group data was assessed with one- or two-way analyses of variance followed by Bonferroni post-hoc tests when indicated by significant main effects or interactions; two-group data were analysed with two-tailed Student's *t*-tests. All correlations were analysed with Pearson's correlation test. All procedures were approved by the Gallo Center Institutional Animal Care and Use Committee and were in accordance with National Institutes of Health guidelines.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 9 November 2007; accepted 4 April 2008.
Published online 11 May 2008.

1. LeDoux, J. The emotional brain, fear, and the amygdala. *Cell. Mol. Neurobiol.* **23**, 727–738 (2003).
2. Davis, M. in *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction* (ed. Aggleton, J. P.) 255–306 (Wiley, Chichester, UK, 1992).
3. Rosenkranz, J. A. & Grace, A. A. Dopamine-mediated modulation of odour-evoked amygdala potentials during pavlovian conditioning. *Nature* **417**, 282–287 (2002).
4. Maren, S. & Quirk, G. J. Neuronal signalling of fear memory. *Nature Rev. Neurosci.* **5**, 844–852 (2004).
5. Cador, M., Robbins, T. W. & Everitt, B. J. Involvement of the amygdala in stimulus–reward associations: interaction with the ventral striatum. *Neuroscience* **30**, 77–86 (1989).
6. Cardinal, R. N., Parkinson, J. A., Hall, J. & Everitt, B. J. Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. *Neurosci. Biobehav. Rev.* **26**, 321–352 (2002).
7. Balleine, B. W. & Killcross, S. Parallel incentive processing: an integrated view of amygdala function. *Trends Neurosci.* **29**, 272–279 (2006).
8. Schoenbaum, G., Chiba, A. A. & Gallagher, M. Neural encoding in orbitofrontal cortex and basolateral amygdala during olfactory discrimination learning. *J. Neurosci.* **19**, 1876–1884 (1999).
9. Uwano, T., Nishijo, H., Ono, T. & Tamura, R. Neuronal responsiveness to various sensory stimuli, and associative learning in the rat amygdala. *Neuroscience* **68**, 339–361 (1995).

10. Tye, K. M. & Janak, P. H. Amygdala neurons differentially encode motivation and reinforcement. *J. Neurosci.* **27,** 3937–3945 (2007).
11. Paton, J. J., Belova, M. A., Morrison, S. E. & Salzman, C. D. The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature* **439,** 865–870 (2006).
12. Doron, N. N. & Ledoux, J. E. Organization of projections to the lateral amygdala from auditory and visual areas of the thalamus in the rat. *J. Comp. Neurol.* **412,** 383–409 (1999).
13. Azuma, S., Yamamoto, T. & Kawamura, Y. Studies on gustatory responses of amygdaloid neurons in rats. *Exp. Brain Res.* **56,** 12–22 (1984).
14. Nakashima, M. *et al.* An anterograde and retrograde tract-tracing study on the projections from the thalamic gustatory area in the rat: distribution of neurons projecting to the insular cortex and amygdaloid complex. *Neurosci. Res.* **36,** 297–309 (2000).
15. McDonald, A. J. Cortical pathways to the mammalian amygdala. *Prog. Neurobiol.* **55,** 257–332 (1998).
16. Ungless, M. A., Whistler, J. L., Malenka, R. C. & Bonci, A. Single cocaine exposure *in vivo* induces long-term potentiation in dopamine neurons. *Nature* **411,** 583–587 (2001).
17. Perkel, D. J. & Nicoll, R. A. Evidence for all-or-none regulation of neurotransmitter release: implications for long-term potentiation. *J. Physiol. (Lond.)* **471,** 481–500 (1993).
18. Malenka, R. C. & Nicoll, R. A. Long-term potentiation—a decade of progress? *Science* **285,** 1870–1874 (1999).
19. Hess, G., Kuhnt, U. & Voronin, L. L. Quantal analysis of paired-pulse facilitation in guinea pig hippocampal slices. *Neurosci. Lett.* **77,** 187–192 (1987).
20. Shin, R. M., Tsvetkov, E. & Bolshakov, V. Y. Spatiotemporal asymmetry of associative synaptic plasticity in fear conditioning pathways. *Neuron* **52,** 883–896 (2006).
21. Humeau, Y., Shaban, H., Bissiere, S. & Luthi, A. Presynaptic induction of heterosynaptic associative plasticity in the mammalian brain. *Nature* **426,** 841–845 (2003).
22. Burns, L. H., Everitt, B. J. & Robbins, T. W. Intra-amygdala infusion of the *N*-methyl-D-aspartate receptor antagonist AP5 impairs acquisition but not performance of discriminated approach to an appetitive CS. *Behav. Neural Biol.* **61,** 242–250 (1994).
23. Baldwin, A. E., Holahan, M. R., Sadeghian, K. & Kelley, A. E. *N*-methyl-D-aspartate receptor-dependent plasticity within a distributed corticostriatal network mediates appetitive instrumental learning. *Behav. Neurosci.* **114,** 84–98 (2000).
24. Rosenkranz, J. A., Moore, H. & Grace, A. A. The prefrontal cortex regulates lateral amygdala neuronal plasticity and responses to previously conditioned stimuli. *J. Neurosci.* **23,** 11054–11064 (2003).
25. Samson, R. D. & Pare, D. A spatially structured network of inhibitory and excitatory connections directs impulse traffic within the lateral amygdala. *Neuroscience* **141,** 1599–1609 (2006).
26. McKernan, M. G. & Shinnick-Gallagher, P. Fear conditioning induces a lasting potentiation of synaptic currents *in vitro*. *Nature* **390,** 607–611 (1997).
27. Rumpel, S., LeDoux, J., Zador, A. & Malinow, R. Postsynaptic receptor trafficking underlying a form of associative learning. *Science* **308,** 83–88 (2005).
28. Tsvetkov, E., Carlezon, W. A., Benes, F. M., Kandel, E. R. & Bolshakov, V. Y. Fear conditioning occludes LTP-induced presynaptic enhancement of synaptic transmission in the cortical pathway to the lateral amygdala. *Neuron* **34,** 289–300 (2002).
29. Quirk, G. J., Armony, J. L. & LeDoux, J. E. Fear conditioning enhances different temporal components of tone-evoked spike trains in auditory cortex and lateral amygdala. *Neuron* **19,** 613–624 (1997).
30. McGaugh, J. L. Memory consolidation and the amygdala: a systems perspective. *Trends Neurosci.* **25,** 456–461 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank H. L. Fields, R. A. Nicoll, A. J. Doupe, B. T. Chen, M. J. Wanat and F. W. Hopf for critical comments; W. W. Schairer, J. J. Cone and L. D. Tye for technical assistance; and T. M. Gill and A. D. Milstein for discussion and technical advice. This study was supported by the State of California for Medical Research on Alcohol and Substance Abuse through the University of California at San Francisco (P.H.J. and A.B.), National Institutes of Health grant RO1DA115096 (A.B.) and a National Science Foundation Graduate Research Fellowship (K.M.T.).

Author Contributions K.M.T. performed the experiments and analyzed the data, with assistance and training in whole-cell recording from G.D.S., who performed pilot mEPSC experiments. B.R. performed cannula surgeries and trained K.M.T. in microinjection techniques. A.B. and P.H.J. provided mentorship and resources. K.M.T., G.D.S., A.B. and P.H.J. contributed to study design, results analysis, interpretation and manuscript writing.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.J. (pjanak@gallo.ucsf.edu).

METHODS

Behavioural training. Nose-poke responses were reinforced on about 50% of trials with a subsequent (onset 50 ms after nose-poke) light–tone cue: 3-kHz tone at 80 dB and illumination of two 5-s stimulus lights. At 2 s after the nose-poke, 0.1 ml of 15% sucrose was delivered to a port adjacent to the nose-poke operandum over 3 s. Additional rewards could not be earned until the previous sucrose reward had been consumed (as determined by port entries). Therefore, to maintain contingency between the cue and the reward, whenever sucrose was present in the reward port, all nose-poke responses were paired with the cue. Hence, early learning sessions tended to result in higher percentages of nose-pokes that were presented together with the cue (mean 56%, maximum 70%, minimum 39%), and in higher numbers of cue presentations than sucrose deliveries. Behavioural indices were calculated for each session. The behavioural indices, namely task efficiency and task accuracy, measured distinct aspects of reward-learning success. Task efficiency (rewards earned divided by cues presented) measured the strength of the cue–reward association, because each cue presentation signalled an opportunity for the rats to collect sucrose at the adjacent reward port. Task accuracy (the difference between the number of correct and incorrect trials divided by total port entries) measured each rat's ability to predict accurately when sucrose would be present in the reward port.

***In vivo* electrophysiology.** Rats were bilaterally implanted with fixed eight-wire electrode arrays (NeuroBiological Laboratories) in the vLA (anteroposterior (AP), −2.8 to −3.3 mm; mediolateral (ML), ±5.0 mm; dorsoventral (DV), 7.2 mm) for chronic neural recording during learning in a custom operant conditioning chamber (MedAssociates) as in ref. 10. Neural activity was recorded, and unit discrimination was performed, with multichannel spike acquisition and sorting software (Plexon Inc.). Responses of single units were deemed statistically significant if the firing rate within one or more 100-ms bins in the response window (0–0.5 s after cue onset) was significantly different ($P < 0.01$) from a 0.5-s baseline epoch (−2 to −1.5 s) using a Wilcoxon signed-rank test. To determine whether single units developed a within-session cue response, the Mann–Whitney *U*-test was used to compare trials from the pre-acquisition and post-acquisition epochs for five 100-ms bins in the first 500 ms response window following cue onset. For all comparisons between pre-acquisition and post-acquisition, the number of trials chosen was determined by the epoch (pre-acquisition versus post-acquisition) with the fewest trials, and an equivalent number of trials was randomly selected from the other epoch. For the peri-event surface plot (Fig. 1c), spike counts for each unit were converted to *Z*-scores by $[(FR_i - FR_{mb})/SD_b]$, where FR_i is the firing rate in the *i*th bin of the peri-event period, FR_{mb} is the mean and SD_b is the standard deviation of the firing rate of a baseline period across the session (between 1.5 and 2 s before the event). Each unit was then smoothed by averaging each trial with its neighbouring trials (±1) and units were averaged to construct a peri-event surface plot (MATLAB; Mathworks) showing the activity of all recorded units ($n = 13$) from one rat as the task was acquired. For the population peri-event histogram shown in Fig. 1e, *Z*-scores were calculated in 50-ms bins for each individual neuron relative to baseline and response periods per trial, and averaged to reveal the population response for sessions 1 and 3.

***Ex vivo* electrophysiology.** About 30 min after session end, rats were anaesthetized with 40 mg kg^{-1} pentobarbital and perfused transcardially with 30 ml of modified aCSF (at about 1 °C) for perfusion containing (in mM): 225 sucrose, 126 NaCl, 2.5 KCl, 1.0 NaH_2PO_4, 4.9 $MgCl_2$, 0.1 $CaCl_2$, 26.2 $NaHCO_3$, 1.25 glucose, 3 kynurenic acid. Coronal sections containing the LA (320 μm) were collected in a holding chamber (superfusion solution, saturated with 95% O_2 and 5% CO_2, containing (in mM): 126 NaCl, 2.5 KCl, 1.0 NaH_2PO_4, 1.3 $MgCl_2$, 2.4 $CaCl_2$, 26.2 $NaHCO_3$, 11 glucose, 1 ascorbic acid at 32–34 °C)) to recover for about 1 h before recording with the same superfusion solution without ascorbic acid but with 0.1 mM picrotoxin. Recordings were made from visually identified pyramidal neurons in the ventral aspect of the LA. Recording electrodes (2.8–4.0 MΩ) were filled with (in mM): 120 caesium methansulphonate, 20 HEPES, 0.4 EGTA, 2.8 NaCl, 5 tetraethylammonium chloride, 2.5 MgATP, 0.25 NaGTP (pH 7.25–7.4; 280–290 milli-osM). Series resistance (10–20 MΩ) and input resistance were monitored online. EPSCs were filtered at 2 kHz and collected with custom scripts written in IgorPro software (Wavemetrics). The AMPAR/NMDAR ratio was calculated by averaging 20–30 EPSCs at +40 mV before and after application of the NMDAR blocker AP5 (50 μM) for 5 min. NMDAR responses were calculated by subtracting the average response in the presence of AP5 from that seen in its absence. Similar to previous studies[20,27,28], electrical stimulation was applied to the internal capsule to evoke EPSCs in LA neurons from thalamic afferents[12], and the external capsule to evoke EPSCs from cortical afferents[15]. In each rat from which a thalamic AMPAR/NMDAR ratio was recorded, a cortical AMPAR/NMDAR ratio was also recorded. mEPSC traces were filtered at 1 kHz, collected with Clampex (Molecular Devices) and analysed with Mini Analysis Program (Synaptosoft). AMPAR mEPSCs were recorded in cells voltage-clamped at −70 mV and in the continual presence of lidocaine (500 μM) over 5 min; 300 events were analysed per cell (the detection criterion was set at 7 pA). Behavioural performance was not calculated until after whole-cell recordings had been analysed.

Intra-LA infusions. Rats were implanted with cannulae just dorsal to the LA (AP, −2.8 to −3.3 mm; ML, ± 5.0 mm; DV, 7.0 mm). One week later, rats received sham infusions of aCSF 24 h before the training session; 10–15 min before the training session, aCSF or AP5 (0.4 μl aCSF per side or 3 μg of AP5 per 0.4 μl per side; 0.1 μl min^{-1}) was infused bilaterally. After training, brains were prepared for whole-cell recordings as above, after careful removal of the cannula headstage. Cannula placements were revealed during the slice recording session with an upright microscope under infrared illumination (Supplementary Fig. 6). An additional group received cannulae in the central nucleus of the amygdala (AP, −1.8 to 2.3 mm; ML, ±4.6 mm; DV, 7.0 mm) for infusion of AP5.

LETTERS

Crystal structures of oseltamivir-resistant influenza virus neuraminidase mutants

Patrick J. Collins[1], Lesley F. Haire[1], Yi Pu Lin[1], Junfeng Liu[1], Rupert J. Russell[2], Philip A. Walker[1], John J. Skehel[1], Stephen R. Martin[1], Alan J. Hay[1] & Steven J. Gamblin[1]

The potential impact of pandemic influenza makes effective measures to limit the spread and morbidity of virus infection a public health priority. Antiviral drugs are seen as essential requirements for control of initial influenza outbreaks caused by a new virus, and in pre-pandemic plans there is a heavy reliance on drug stockpiles. The principal target for these drugs is a virus surface glycoprotein, neuraminidase, which facilitates the release of nascent virus and thus the spread of infection. Oseltamivir (Tamiflu) and zanamivir (Relenza) are two currently used neuraminidase inhibitors that were developed using knowledge of the enzyme structure[1,2]. It has been proposed that the closer such inhibitors resemble the natural substrate, the less likely they are to select drug-resistant mutant viruses that retain viability[3]. However, there have been reports of drug-resistant mutant selection *in vitro*[4] and from infected humans[5,6]. We report here the enzymatic properties and crystal structures of neuraminidase mutants from H5N1-infected patients that explain the molecular basis of resistance. Our results show that these mutants are resistant to oseltamivir but still strongly inhibited by zanamivir owing to an altered hydrophobic pocket in the active site of the enzyme required for oseltamivir binding. Together with recent reports of the viability and pathogenesis of H5N1 (ref. 7) and H1N1 (ref. 8) viruses with neuraminidases carrying these mutations, our results indicate that it would be prudent for pandemic stockpiles of oseltamivir to be augmented by additional antiviral drugs, including zanamivir.

Influenza neuraminidase (NA) functions in virus infection to remove sialic acid from cell-surface receptors so that newly made viruses are released and able to spread to uninfected cells[9]. To inhibit virus propagation NA inhibitors were developed, informed in part by the crystal structures of several NAs belonging to one of the two genetically defined NA groups of influenza A[10,11], but with the objective that they should be active against all influenza viruses[12]. The outcome of this work includes the licensing of the drugs oseltamivir (Tamiflu)[1] and zanamivir (Relenza)[2], which are active against NAs from group 1 and 2 influenza A as well as influenza B viruses[5]. From this work it was proposed that inhibitors that closely resemble the natural sialic acid substrate of NA are unlikely to select drug-resistant mutants that retain normal enzyme activity[3]. Thus, viruses that are resistant to such drugs ought to be less viable[13–15]. In the present study we have determined binding and inhibitory parameters for oseltamivir and zanamivir interacting with wild-type and three mutant H5N1 neuraminidases. The mutations involved are: (1) His274Tyr, the principal mutation isolated in association with oseltamivir treatment that is specific to the N1 group[16] and that has recently been shown to be present in substantial numbers of H1N1 viruses isolated from humans[8]; (2) Asn294Ser, which has been identified in viruses containing either N1 or N2 NAs, isolated from patients treated with oseltamivir[17,18]; and (3) a variant of N1 in which the generally conserved tyrosine residue at position 252 was replaced by histidine. This substitution was not associated with drug treatment and was found only in one clade of H5N1 viruses isolated from infected humans in Vietnam and Cambodia in 2004 and 2005 (ref. 19).

We used a fluorescent substrate (see Methods and Supplementary Information) to measure enzyme activity (V_m) and Michaelis constant (K_m) as well as inhibitory constants (K_I) for the different NAs. The values for enzyme activity (V_m) given in Table 1 for wild-type and mutant NAs are similar, and there is at most a ninefold difference in the K_m values. These results are similar to those previously reported[20,21] and support the notion that substitution of amino acids adjacent to the active site, that do not interact with the substrate, can be accommodated without significantly affecting NA activity[22]. Inspection of the inhibitory constants (K_I) in Table 1 shows that none of the mutations has a major impact on the effectiveness of zanamivir. However, two of them substantially reduce the inhibition caused by oseltamivir: the His274Tyr substitution by a factor of 265 and the Asn294Ser by a factor of 81. In contrast, the Tyr252His substitution leads to a tenfold increase in sensitivity to this inhibitor.

We compared the kinetics of oseltamivir and zanamivir binding to wild-type NA (see Methods) to show that the approach to equilibrium is approximately three times faster with oseltamivir and to

Table 1 | Activity, binding and kinetic parameters for N1 neuraminidases

NA type	V_m relative to wild type	K_m (μM)	Oseltamivir relative K_I*	Zanamivir relative K_I†	k_{on} ($\mu M^{-1} s^{-1}$) oseltamivir	k_{off} (s^{-1}) oseltamivir ($\times 10^4$)	k_{on} ($\mu M^{-1} s^{-1}$) zanamivir	k_{off} (s^{-1}) zanamivir ($\times 10^4$)
Wild type	1.0	6.3	1.0	1.0	2.52 (0.21)	8.1 (1.2)	0.95 (0.08)	0.95 (0.13)
His274Tyr	0.8	27.0	265	1.9	0.24 (0.06)	180 (30)‡	0.35 (0.02)	0.67 (0.08)
Asn294Ser	1.15	53.0	81	7.2	1.1 (0.18)	235 (40)‡	0.52 (0.04)	3.7 (0.6)
Tyr252His	0.94	7.5	0.1	1.2	3.9 (0.15)	1.25 (0.13)	1.38 (0.15)	1.66 (0.33)

K_m values are from three determinations; K_I values from at least six measurements. Values in parentheses represent the standard deviations obtained from linear least squares fits to k_{obs} values as a function of substrate and inhibitor concentrations, as shown in Supplementary Information. k_{on} and k_{off} are the association and dissociation rate constants, respectively.
* Oseltamivir relative K_I is K_I(mutant)/K_I(wild type), where wild type = 0.32 nM.
† Zanamivir relative K_I is K_I(mutant)/K_I(wild-type), where wild type = 0.1 nM.
‡ Directly determined.

[1]MRC-National Institute for Medical Research, Mill Hill, London NW7 1AA, UK. [2]Interdisciplinary Centre for Human and Avian Influenza Research, School of Biology, University of St Andrews, Fife KY16 9ST, UK.

confirm that zanamivir is the slightly more potent inhibitor (Fig. 1a, b and Table 1). Our kinetic data also show that between the wild-type and mutant enzymes there are, as expected, much greater differences for the rates of association and dissociation with oseltamivir than with zanamivir. The reduced affinity of the His274Tyr mutant NA for oseltamivir arises from a tenfold poorer association rate constant and a 25-fold enhanced dissociation rate constant. For the Asn294Ser mutant the reduced affinity is largely accounted for by a 35-fold increase in off-rate for the inhibitor whereas the Tyr252His variant is more sensitive to oseltamivir mostly because of a 6.5-fold decrease in the off-rate. Notably, there is close agreement between the inhibitory constants and the ratio of association and dissociation rates measured kinetically, giving us confidence in the robustness of the data.

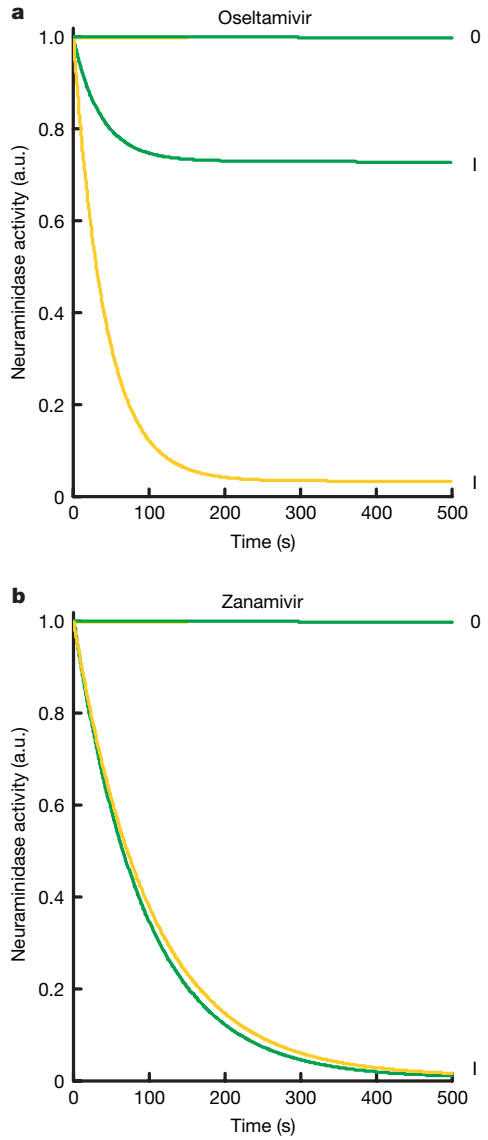


Figure 1 | Neuraminidase activity monitored using a fluorescent assay. NA activity for wild type (yellow) and His274Tyr (green) mutant proteins in the absence (labelled 0) and presence (labelled I) of 85 nM inhibitor at 50 μM substrate. **a**, **b**, Effect of oseltamivir (**a**) and zanamivir (**b**). For the His274Tyr mutant the approach to equilibrium occurs at a similar rate for the two inhibitors even though oseltamivir is the much poorer inhibitor. At the oseltamivir concentration used the reduced contribution of the on-rate constant to the observed rate is almost exactly compensated for by the increased contribution from the off-rate constant. Although oseltamivir has sometimes been referred to as a slow-binding inhibitor[21,30], the association rate constants that we determined are within the range frequently observed for the interaction of small molecules with proteins. It is the use both here and elsewhere of very low inhibitor concentrations that result in a slow approach to equilibrium, not the kinetics of binding.

We used X-ray crystallography to understand the structural basis for the poorer binding of oseltamivir to the His274Tyr and Asn294Ser mutant NAs relative to wild type (Fig. 2a, b). Crystals of the His274Tyr mutant NA in complex with both inhibitors diffracted to high resolution and produced clearly defined electron density for the bound oseltamivir (Fig. 2c) and zanamivir (Fig. 2d). Relevant crystallographic statistics are given in Supplementary Table 1. With respect to binding to wild-type NA the key difference between oseltamivir and zanamivir is that oseltamivir has a hydrophobic pentyloxy substituent at the C6 position rather than a polar glycerol group in zanamivir (as is also the case for the sialic acid substrate, Fig. 2b). Binding of oseltamivir to wild-type NA involves a conformational change in the side chain of Glu 276 relative to the ligand-free enzyme[1,3,23]. In its new conformation, the carboxyl group of Glu 276 is oriented away from the hydrophobic pentyloxy group of oseltamivir, the latter making hydrophobic contact with the Cβ methylene of Glu 276 (Fig. 2c, yellow coloured chain). Binding of zanamivir, like sialic acid[24], involves hydrogen-bond formation between the carboxyl group of Glu 276 and the 8- and 9-hydroxyl groups of the glycerol moiety of the inhibitor and requires no change in side-chain conformation (Fig. 2d, yellow coloured chain). The structure of the His274Tyr–oseltamivir complex (Fig. 2c) shows that substitution by the bulkier tyrosine residue pushes the carboxyl group of Glu 276 2 Å farther into the binding site. In this position the charged group disrupts the otherwise hydrophobic pocket that normally accommodates the pentyloxy substituent of oseltamivir and causes a change in the conformation of the inhibitor such that its C9 and C91 carbons move about 2.5 Å from their wild-type NA-bound position (Fig. 2c). This result is similar to the observations on inhibitor binding to the Arg292Lys mutation of N9 (ref. 3). In contrast, the structure of the His274Tyr–zanamivir complex (Fig. 2d) shows how the tyrosine residue is accommodated by a small movement in the side chain of Glu 276. Moreover, the adjustment is made without disrupting the hydrogen bonds made between zanamivir and Glu 276 in wild-type NA.

The two structures thus provide a direct explanation for the reduction in the binding affinity of the His274Tyr mutant for oseltamivir without significantly altering the binding of zanamivir. The structures also confirm our earlier suggestion that the substitution of tyrosine for histidine at position 274 would not be accommodated in group 1 (N1, N4, N5 and N8) NAs without disrupting the oseltamivir-binding site, because of the presence of a bulky and conserved tyrosine at position 252 beneath residue 274 (Fig. 2a). In contrast, group 2 NAs (N2, N3, N6, N7 and N9) have the smaller threonine residue at position 252 and can accommodate the His274Tyr substitution without changing the oseltamivir-binding site[25]. Our binding and structural studies also provide support for the notion that zanamivir, having the same glycerol moiety at C6 as sialic acid (Fig. 2b), is less likely to encounter mutations in this region of the active site that weaken its binding without also weakening the binding of the natural substrate.

In N1 NAs residue 252 is generally conserved as tyrosine. However, in a number of H5N1 viruses isolated from humans in Vietnam in 2004–05 histidine was present at position 252 (ref. 19). As our thermodynamic and kinetics data show, this substitution results in an enzyme about 10 times more sensitive to oseltamivir but essentially unchanged with respect to zanamivir binding. We suspect that we have been unable to crystallize this Tyr252His mutant because of a more flexible structure arising from the loss of hydrogen bonds of the tyrosine side chain with both main-chain groups and the side chain of His 274. Our data indicating that the Tyr252His mutation affects oseltamivir but not zanamivir binding suggests that it is the loss of the Tyr 252 to His 274 hydrogen bond that favours the re-orientation of Glu 276 required for oseltamivir binding, thus accounting for the

tenfold enhancement of oseltamivir binding by this mutant NA without altering its affinity for zanamivir.

We have also determined the structure of the complex of Asn294Ser with oseltamivir (Fig. 2e). This shows smaller changes in the position of the mutant-bound oseltamivir and adjacent side chains, relative to wild-type, than those seen for oseltamivir bound to the His274Tyr mutant (Fig. 2c, f). The Ser 294 side chain in the mutant is oriented so that its polar hydroxyl group forms a hydrogen bond with the carboxylate of Glu 276. By contrast, the side chain of Asn 294 in the wild-type NA is directed in approximately the opposite direction, towards Tyr 347. The loss of the asparagine side chain at position 294 enables the main-chain carbonyl of Tyr 347 to flip out, from its position in the wild-type, so that it no longer coordinates to the bound calcium ion (blue sphere in Fig. 2e). These changes lead to a less well-ordered conformation for the side chain of Tyr 347 and, presumably, a weaker hydrogen bond interaction with the carboxylate of oseltamivir (and by inference with the equivalent carboxylate in zanamivir and sialic acid). This interpretation of the Asn294Ser–oseltamivir structure is consistent with the observation (Table 1) that this mutant has an approximately sevenfold weaker binding for zanamivir (and about an eightfold higher K_m for its substrate). We suspect that the 81-fold weaker binding of this mutant to oseltamivir results from a combination of the effect on Tyr 347 with the effect of the substituted serine residue hydrogen bonding to the side chain of Glu 276. The hydrogen bond stabilizes the serine side-chain conformation so that the polar hydroxyl group is brought into what is otherwise a hydrophobic binding site for oseltamivir in wild-type NA. Such a perturbation of the hydrophobic make-up of the binding site, which would be expected to affect the binding of oseltamivir but not zanamivir, is consistent with our

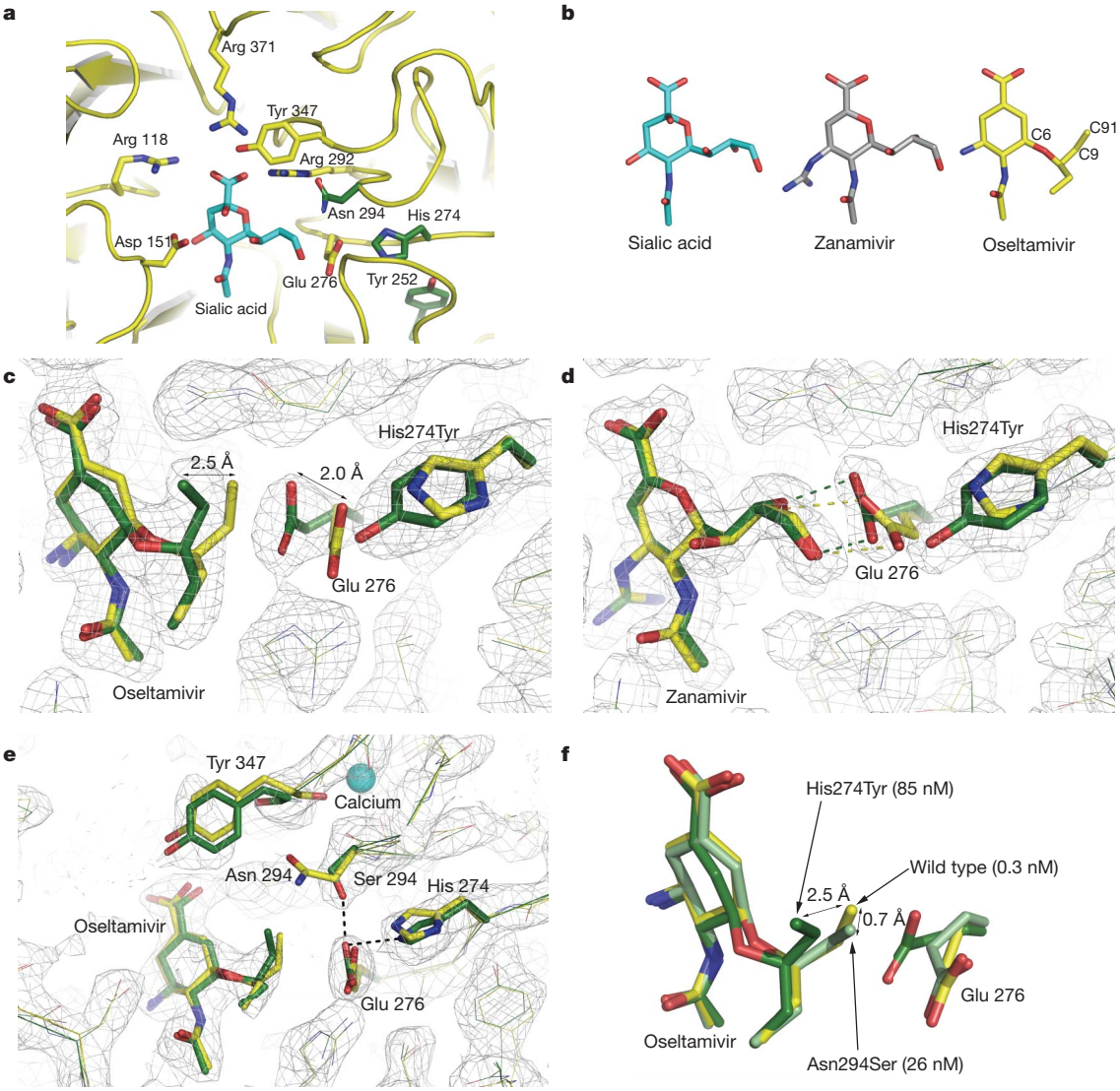


Figure 2 | Structure of N1 neuraminidase complexes. **a**, Sialic acid (coloured blue) docked into the active site of wild-type N1 NA (ribbons coloured yellow) from superposition of the sialic acid complex of N2 (ref. 24) (Protein Data Bank code 2BAT). The positions of some key binding residues are shown with carbons coloured yellow, nitrogens blue and oxygens in red. The side chains of the three mutants examined in this work are shown with their carbons coloured in green (shown as the wild-type residues). **b**, The structures of sialic acid (carbons coloured blue), zanamivir (carbons coloured grey) and oseltamivir (carbons coloured yellow) are shown in similar orientations with selected carbon atoms numbered. **c**–**e**, The overlaid structures of the active sites of wild-type (yellow) and mutant N1 NAs (green) are shown with bound inhibitors coloured similarly; relevant portions of electron density maps are also shown. **c**, His274Tyr in complex with oseltamivir. **d**, His274Tyr in complex with zanamivir. **e**, Asn294Ser in complex with oseltamivir. Dashed lines indicate selected hydrogen bonds. Electron density maps were calculated from $2F_o - F_c$, α_{calc} coefficients and are contoured at 1.2σ. **f**, The conformation of oseltamivir and Glu 276 from three complexes is shown after superposition using protein atoms only; the carbon atoms of the inhibitor from the wild-type complex are coloured yellow, the His274Tyr in dark green and the Asn294Ser in light green. The affinities (K_I) of oseltamivir for the three NAs are given in parentheses.

measurements of the relative affinities of this mutant for the two inhibitors.

There are currently two groups of anti-influenza drugs that target either the M2 proton channel or NA. Mutations in M2 have caused resistance among both H3N2 and H1N1 current seasonal influenza viruses and a large proportion of the avian H5N1 viruses to the drugs amantadine and rimantadine[26], which block the channel, placing extra dependence on the NA-inhibitor drugs oseltamivir and zanamivir. Of these, oseltamivir has been used more frequently than zanamivir but there is little information for either drug on the rate of emergence of resistant variants in the case of the H5N1 viruses. Nevertheless, mutant viruses have been isolated from fatal H5N1 infections and among these the His274Tyr mutation described here is prominent[6]. Our results explain the structural basis for the resistance of this mutant to oseltamivir and for its sensitivity to zanamivir. While our work was in progress, the biological properties of H5N1 viruses carrying the His274Tyr mutation were described from studies of infected mice[7]. Previously, in ferrets, both wild-type and mutant viruses were viable, with the mutant producing only about ten times less virus than the wild-type at the peak of infection[18]. In the more recent studies of infected mice, both viability and pathogenicity were also closely similar for wild-type and mutant viruses[7]. In contrast with previous reports[13,14], the viability of viruses containing the His274Tyr mutant NA has also been demonstrated recently by the isolation of numerous oseltamivir-resistant H1N1 viruses from outbreaks of human infections in Europe[8]. Like the mutant NAs detailed here, these enzymes are both resistant to oseltamivir and sensitive to zanamivir even though the patients from whom the mutant viruses were isolated had no known exposure to oseltamivir. Evidently, if the His274Tyr mutation also becomes common in avian H5N1, the effectiveness of oseltamivir would be limited and the value of stockpiles compromised. Considered together with the success of HAART (highly active antiretroviral therapy) against HIV[27,28], it would be prudent to re-assess the suitability of single-drug, pre-pandemic stockpiles and to develop effective drug combination treatments.

METHODS SUMMARY

Site-directed mutagenesis was performed on the N1 neuraminidase gene of A/Vietnam/1203/04 (H5N1). Mutant N1 neuraminidase (NA) was prepared from a WSN-NA (H1N1) recombinant virus containing seven genes from WSN and the mutated N1 neuraminidase gene of H5N1. Recombinant viruses were grown in hens' eggs and the neuraminidase was released from the virus by bromelain digestion, and further purified, as previously described[29]. Purified NA concentrations were determined from their absorption spectra and enzymatic activity was measured using the fluorescent substrate 2′-(4-methylumbelliferyl)-α-D-*N*-acetylneuraminic acid (MUNANA) using a JASCO FP-6300 fluorimeter with excitation and emission wavelengths of 365 and 450 nm. NA activity changes in the presence of inhibitors were monitored as the first derivative of the fluorescence change. Crystals were obtained by vapour diffusion using 15% PEG 3350 as precipitant. Diffraction data were collected at 100K and processed using Denzo and Scalepack. The structures were solved by molecular replacement and refined using Refmac5 or PHENIX with manual model building using O or Coot.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 14 February; accepted 1 April 2008.
Published online 14 May 2008.

1. Kim, C. U. *et al.* Influenza neuraminidase inhibitors possessing a novel hydrophobic interaction in the enzyme active site: design, synthesis, and structural analysis of carbocyclic sialic acid analogues with potent anti-influenza activity. *J. Am. Chem. Soc.* **119,** 681–690 (1997).
2. von Itzstein, M. *et al.* Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* **363,** 418–423 (1993).
3. Varghese, J. N. *et al.* Drug design against a shifting target: a structural basis for resistance to inhibitors in a variant of influenza virus neuraminidase. *Structure* **6,** 735–746 (1998).
4. McKimm-Breschkin, J. L. Resistance of influenza viruses to neuraminidase inhibitors–a review. *Antiviral Res.* **47,** 1–17 (2000).
5. Gubareva, L. V., Webster, R. G. & Hayden, F. G. Comparison of the activities of zanamivir, oseltamivir, and RWJ-270201 against clinical isolates of influenza virus and neuraminidase inhibitor-resistant variants. *Antimicrob. Agents Chemother.* **45,** 3403–3408 (2001).
6. de Jong, M. D. *et al.* Oseltamivir resistance during treatment of influenza A (H5N1) infection. *N. Engl. J. Med.* **353,** 2667–2672 (2005).
7. Yen, H. L. *et al.* Neuraminidase inhibitor-resistant recombinant A/Vietnam/1203/04 (H5N1) influenza viruses retain their replication efficiency and pathogenicity *in vitro* and *in vivo*. *J. Virol.* **81,** 12418–12426 (2007).
8. Lackenby, A. *et al.* Emergence of resistance to oseltamivir among influenza A(H1N1) viruses in Europe. *Euro Surveill.* **13** (2008).
9. Murphy, B. R. & Webster, R. G. in *Fields Virology* 3rd edn (eds Fields, D. B. N., Knipe, M. & Howley, P. M.) 1397–1445 (Lippincott-Raven, Philadelphia, 1996).
10. Varghese, J. N., Laver, W. G. & Colman, P. M. Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9 Å resolution. *Nature* **303,** 35–40 (1983).
11. Baker, A. T., Varghese, J. N., Laver, W. G., Air, G. M. & Colman, P. M. Three-dimensional structure of neuraminidase of subtype N9 from an avian influenza virus. *Proteins* **2,** 111–117 (1987).
12. Burmeister, W. P., Henrissat, B., Bosso, C., Cusack, S. & Ruigrok, R. W. Influenza B virus neuraminidase can synthesize its own inhibitor. *Structure* **1,** 19–26 (1993).
13. Ives, J. *et al.* Anti-viral drug resistance: An oseltamivir treatment-selected influenza A/N2 virus with a R292K mutation in the neuraminidase gene has reduced infectivity *in vivo*. *J. Clin. Virol.* **18,** 251–269 (2000).
14. Ives, J. A. L. *et al.* The H274Y mutation in the influenza A/H1N1 neuraminidase active site following oseltamivir phosphate treatment leave virus severely compromised both *in vitro* and *in vivo*. *Antiviral Res.* **55,** 307–317 (2002).
15. Herlocher, M. L. *et al.* Influenza viruses resistant to the antiviral drug oseltamivir: transmission studies in ferrets. *J. Infect. Dis.* **190,** 1627–1630 (2004).
16. Gubareva, L. V., Kaiser, L., Matrosovich, M. N., Soo-Hoo, Y. & Hayden, F. G. Selection of influenza virus mutants in experimentally infected volunteers treated with oseltamivir. *J. Infect. Dis.* **183,** 523–531 (2001).
17. Kiso, M. *et al.* Resistant influenza A viruses in children treated with oseltamivir: descriptive study. *Lancet* **364,** 759–765 (2004).
18. Le, Q. M. *et al.* Avian flu: isolation of drug-resistant H5N1 virus. *Nature* **437,** 1108 (2005).
19. McKimm-Breschkin, J. L., Selleck, P. W., Usman, T. B. & Johnson, M. A. Reduced sensitivity of influenza A to oseltamivir. *Emerg. Infect. Dis.* **13,** 1354–1357 (2007).
20. Rameix-Welti, M. A. *et al.* Natural variation can significantly alter the sensitivity of influenza A (H5N1) viruses to oseltamivir. *Antimicrob. Agents Chemother.* **50,** 3809–3815 (2006).
21. Wang, M. Z., Tai, C. Y. & Mendel, D. B. Mechanism by which mutations at His274 alter sensitivity of influenza a virus N1 neuraminidase to oseltamivir carboxylate and zanamivir. *Antimicrob. Agents Chemother.* **46,** 3809–3816 (2002).
22. Yen, H. L. *et al.* Neuraminidase inhibitor-resistant influenza viruses may differ substantially in fitness and transmissibility. *Antimicrob. Agents Chemother.* **49,** 4075–4084 (2005).
23. Smith, B. J. *et al.* Structural studies of the resistance of influenza virus neuramindase to inhibitors. *J. Med. Chem.* **45,** 2207–2212 (2002).
24. Varghese, J. N., McKimm-Breschkin, J. L., Caldwell, J. B., Kortt, A. A. & Colman, P. M. The structure of the complex between influenza virus neuraminidase and sialic acid, the viral receptor. *Proteins* **14,** 327–332 (1992).
25. Russell, R. J. *et al.* The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design. *Nature* **443,** 45–49 (2006).
26. Cheung, C. L. *et al.* Distribution of amantadine-resistant H5N1 avian influenza variants in Asia. *J. Infect. Dis.* **193,** 1626–1629 (2006).
27. Gulick, R. M. *et al.* Treatment with indinavir, zidovudine, and lamivudine in adults with human immunodeficiency virus infection and prior antiretroviral therapy. *N. Engl. J. Med.* **337,** 734–739 (1997).
28. De Clercq, E. The design of drugs for HIV and HCV. *Nature Rev. Drug Discov.* **6,** 1001–1018 (2007).
29. Ha, Y., Stevens, D. J., Skehel, J. J. & Wiley, D. C. X-ray structures of H5 avian and H9 swine influenza virus hemagglutinins bound to avian and human receptor analogs. *Proc. Natl Acad. Sci. USA* **98,** 11181–11186 (2001).
30. Tai, C. Y. *et al.* Characterization of human influenza virus variants selected *in vitro* in the presence of the neuraminidase inhibitor GS 4071. *Antimicrob. Agents Chemother.* **42,** 3234–3241 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Work at NIMR was funded by the Medical Research Council (UK). This work was also supported in part by the EU FP6 Programme VIRGIL, contract number 503359. R.J.R. thanks the Scottish Funding Council for financial support.

Author Information Structural data have been deposited with the Protein Data Bank with accession codes 3CL0 (His274Tyr–oseltamivir), 3CKZ (His274Tyr–zanamivir) and 3CL2 (Asn294Ser–oseltamivir). Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.J.G. (sgambli@nimr.mrc.ac.uk).

METHODS

Neuraminidase activity measurements. Michaelis–Menten constants (K_m) were determined using standard initial rate measurements with NA concentrations in the range 0.1 to 0.5 nM and MUNANA concentrations in the range 2 to 200 μM. Dissociation constants for enzyme–inhibitor complexes were determined by measuring the reduction in the rate of MUNANA hydrolysis observed in the presence of different concentrations of the inhibitors. The K_I values were determined using equation (1) (Supplementary Information) for data collected at two or more different MUNANA concentrations and three different drug concentrations. The inhibitor concentration was always at least tenfold higher than the NA concentration to allow the approximation $[I] = [I_{total}]$.

The kinetic parameters for the inhibitors were measured in two ways: by adding enzyme to a pre-warmed mixture of MUNANA and inhibitor, or by adding inhibitor to a standard reaction mixture of enzyme and MUNANA approximately 50–200 s after initiation of the reaction. The two approaches gave identical results. In both cases, the exponential approach to the new steady-state rate was analysed using equation (2) (Supplementary Information). Kinetic measurements were made at two or more different MUNANA concentrations and three different drug concentrations. Association (k_{on}) and dissociation (k_{off}) rate constants for inhibitor binding were then determined using equation (3) (Supplementary Information).

Crystals were obtained by the vapour diffusion method from sitting drops dispensed with an Oryx 8 robot (Douglas Instruments). The drops consisted of 100 nl protein (6 mg ml^{-1}) with 1 mM inhibitor (oseltamivir or zanamivir) mixed with 100 nl reservoir solution of 15% PEG 3350, 0.1 M sodium acetate pH 4.6. Crystals were transferred into a cryoprotectant consisting of reservoir solution augmented with 20% (v/v) PEG 400 and inhibitor before flash freezing.

Data sets were recorded on a Raxis4 detector (100 μm scan) mounted on a Rigaku MicroMax 007 HF generator and for Asn294Ser on the MARCCD at Daresbury station 10.1. Diffraction data were integrated using Denzo and scaled with Scalepack[31]. N1 mutant structures were solved by molecular replacement using AmoRe with the wild-type structure (Protein Data Bank code 2HU4) as the initial search model. Standard refinement was carried out with refmac5 (ref. 32) for the two His274Tyr complexes, and PHENIX[33] for the Asn294Ser complex where non-crystallographic symmetry restraints were used. Manual model building was with O[34] and Coot[35]. Figure 2 was created with Pymol (http://pymol.sourceforge.net/).

31. Otwinowski, Z. & Minor, W. in *Data Collection and Processing* (eds Sawyer, L., Isaacs, N. & Bailey, S.) 556–562 (SERC Daresbury Laboratory, Warrington, 1993).
32. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
33. Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58,** 1948–1954 (2002).
34. Jones, T. A., Zhou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47,** 110–119 (1991).
35. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).

Vol 453|26 June 2008|doi:10.1038/nature06942

LETTERS

Assembly reflects evolution of protein complexes

Emmanuel D. Levy[1], Elisabetta Boeri Erba[2], Carol V. Robinson[2] & Sarah A. Teichmann[1]

A homomer is formed by self-interacting copies of a protein unit. This is functionally important[1,2], as in allostery[3–5], and structurally crucial because mis-assembly of homomers is implicated in disease[6,7]. Homomers are widespread, with 50–70% of proteins with a known quaternary state assembling into such structures[8,9]. Despite their prevalence, their role in the evolution of cellular machinery[10,11] and the potential for their use in the design of new molecular machines[12,13], little is known about the mechanisms that drive formation of homomers at the level of evolution and assembly in the cell[9,14]. Here we present an analysis of over 5,000 unique atomic structures and show that the quaternary structure of homomers is conserved in over 70% of protein pairs sharing as little as 30% sequence identity. Where quaternary structure is not conserved among the members of a protein family, a detailed investigation revealed well-defined evolutionary pathways by which proteins transit between different quaternary structure types. Furthermore, we show by perturbing subunit interfaces within complexes and by mass spectrometry analysis[15], that the (dis)assembly pathway mimics the evolutionary pathway. These data represent a molecular analogy to Haeckel's evolutionary paradigm of embryonic development, where an intermediate in the assembly of a complex represents a form that appeared in its own evolutionary history. Our model of self-assembly allows reliable prediction of evolution and assembly of a complex solely from its crystal structure.

Although homomers are central to biology, only anecdotal knowledge exists on their principles of evolution and assembly, and no unifying theory has been proposed. Large increases in structural data in recent years, however, have enabled us to study quaternary structure or spatial arrangement of subunits on a data set of 5,375 unique structures. This data set is ~tenfold greater than any studied previously[16] (Methods). On the basis of this data set, we quantify how often proteins change their quaternary structure, and identify the evolutionary routes taken to do so. Subsequently, as evolution of a complex can be viewed as assembly over a long timescale, we compare evolutionary routes with (dis)assembly routes probed by mass spectrometry.

Homomers can be separated into two main classes of open or closed symmetry. The first class corresponds to open structures that would polymerize to infinity in the absence of limiting factors. Such assemblies (for example, tubulin and actin) are rare in our data set (3%), probably because their innate dynamic character renders them difficult to crystallize. In contrast, closed symmetries are finite in space, and most homomers adopt either cyclic or dihedral symmetry (Fig. 1a), with only a small fraction (1%) having cubic symmetry (not shown). Throughout we denote Cn as a cyclic complex containing n subunits, and Dn as a dihedral complex containing $2n$ subunits.

It has long been observed that smaller complexes are more abundant than larger ones, and even numbers of subunits are favoured over odd numbers[8,9,17]. Here we confirm this observation, with 62% of complexes being dimers. We quantify the different types of symmetries found in homomers and show that the abundance of complexes with even numbers of subunits is due to the prevalence of dihedral complexes. Whenever an option exists for cyclic or dihedral, on average we find an 11-fold preference for dihedral complexes (Fig. 1b). There is an evolutionary explanation for this preference, as the probability that a dihedral complex evolved by random mutation should be higher than the probability for a cyclic complex for at least two reasons: first, at the level of individual interfaces, in dihedral complexes most interfaces are face-to-face (or back-to-back), whereas all interfaces in cyclics are face-to-back (Fig. 1a) and these are less likely to form by random mutation[5,18]; and second, at the level of whole complexes, evolution of dihedral complexes can take place

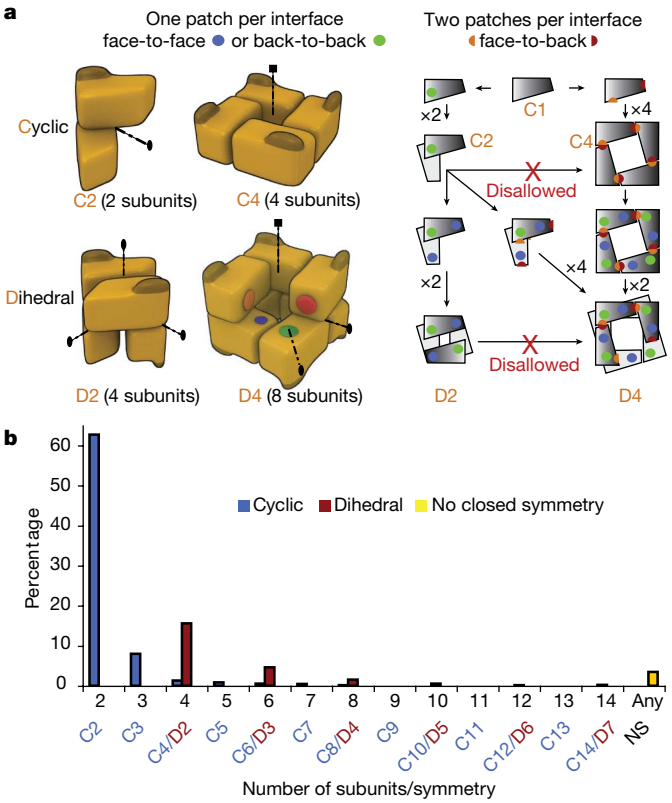


Figure 1 | Abundance and properties of cyclic and dihedral symmetries. **a**, n subunits in a cyclic complex are related by a single n-fold symmetry axis (dotted lines); ellipses and squares represent two- and four-fold axes, respectively. For a monomer to evolve towards a cyclic tetramer (C4), two complementary surfaces have to evolve simultaneously (red and orange patches). For a dihedral tetramer (D2), two different and self-complementary surfaces (green and blue patches) can evolve serially with an intermediate dimer (C2). **b**, The abundance of homomers with cyclic, dihedral, or no symmetry (3.5%). 62.7% are cyclic dimers (C2), 8% are cyclic trimers (C3), and 3.2% have higher order cyclic symmetry (from C4 to C14). Dihedral complexes dominate (22.6%) among complexes with ≥4 subunits.

[1]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK. [2]Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK.

in multiple steps (C1→C2→D2) whereas cyclics must evolve in one step (C1→C4, Fig. 1a).

Notably, dihedral and cyclic symmetries are geometrically related: a complex with D*n* symmetry can be formed from *n* dimers with C2 symmetry or from two *n*-mers with C*n* symmetry[19] (Fig. 1a). If a protein complex has a particular symmetry, we find that homologues are likely to have the same symmetry type. More specifically, for sequence identities >90%, conservation is nearly 100%, whereas in the range of 30–40% sequence identities, conservation is ~70% (Supplementary Fig. 1). Proteins with different degrees of quaternary structure conservation are illustrated in Fig. 2a. Thymidylate synthase always exists as a dimer, adenylyltransferase is a dimer in *Bacillus subtilis* and a hexamer in human (trimer of *B. subtilis* dimers), whereas two phospholipase A2s have geometrically very distant quaternary structures.

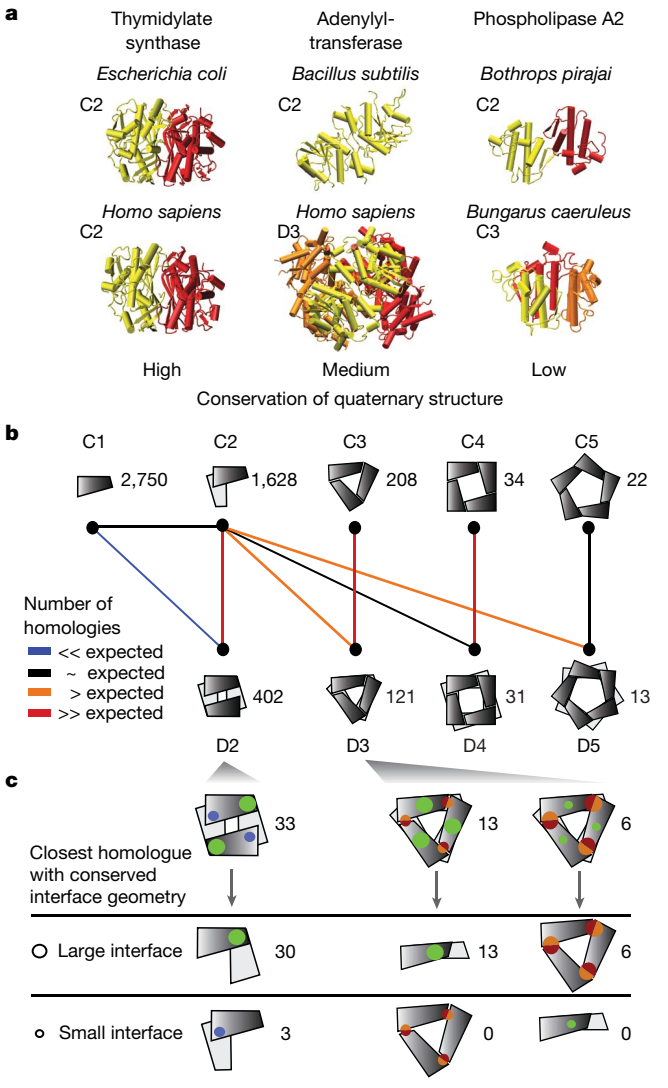


Figure 2 | Routes for homomer evolution. a, Illustrative examples of different levels of quaternary structure conservation (termed high, medium and low). Thymidylate synthases (PDB accessions 1ajm and 1hw3) are both dimers. Adenylyltransferases (PDB accessions 1kam and 1kr2) differ in their number of subunits, with similar dimers (yellow) common to both quaternary structures. The dimer and trimer of venom-toxin phospholipase A2s are not related geometrically and the quaternary structures are therefore less conserved. **b**, Schematic illustration of large-scale analysis of quaternary structures; the number of unique complexes is indicated. Coloured lines indicate the significance in over-representation of shared homologous complexes (Methods). **c**, In 49 out of 52 cases, the largest interface is present in the dimeric or trimeric homologue, illustrated by size of interface patches.

When quaternary structure is not conserved, we speculate that pathways linking geometrically related symmetries represent both evolutionary and assembly routes. For example, a dihedral tetramer (D2) can be described as a dimer of dimers, where a back-to-back dimerization patch forms a first dimer, and a second face-to-face dimerization patch forms the dimer of dimers. This is not true of a cyclic tetramer (C4), where subunits interact in a face-to-back manner, such that two different surface patches are involved in forming an interface (Fig. 1a). Therefore, we expect many more dihedral than cyclic tetramers to share evolutionary relationships with dimers. This is illustrated by the pathway from a dimer to a dihedral tetramer (Fig. 1a) and the disallowed transition from a dimer to a cyclic tetramer.

Following this idea, we looked at evolutionary relationships in terms of sequence similarity between different quaternary structures to unveil the routes most commonly taken to build larger complexes (Fig. 2b). Each quaternary structure is represented schematically with the numbers of proteins of each type. Pairs of quaternary structures are connected according to the statistical significance of the number of evolutionary transitions between them. Most pairs have fewer transitions between them than expected in a random model (Methods) as exemplified by monomers (C1) and dihedral tetramers (D2). Other pairs with insignificant numbers of transitions are shown in Supplementary Fig. 2. We find that cyclic dimers, trimers and tetramers share notable numbers of transitions with their dihedral counterparts, supporting the stepwise evolutionary scenario where homomers with dihedral symmetry evolve through cyclic intermediates (Fig. 1a).

Notably in this stepwise scenario, two evolutionary routes lead to a dihedral complex (D*n*): either from *n* dimers or from two cyclic *n*-mers (Fig. 2b). This raised the question as to whether it was possible to identify which of these two routes was taken by a given dihedral complex. On the basis of energetic considerations (Supplementary Information 1), we propose that a hierarchy of interface sizes exists within dihedral complexes, and that the larger interface is conserved in evolution. To test this hypothesis, we looked for tetramers homologous to a dimer, as well as hexamers homologous to a dimer or trimer. In this data set (Fig. 2c and Supplementary Table 1) we examined whether the interface within the dimer or trimer corresponded to the largest interface in the homologous tetramer or hexamer. Among 33 tetramers and 19 hexamers studied, 49 complexes conserve the larger interface with the dimeric or trimeric homologue, whereas only 3 conserve their smaller interface (Fig. 2c and Supplementary Table 1). This result implies that the evolutionary route of a homomer can be predicted solely from its interface sizes. Our predictions for the evolutionary pathways of D3, D4 and D5 complexes (Supplementary Fig. 3a) have led us to formulate a general model of homomer evolution (Supplementary Fig. 3b).

It is notable that this signature of complex formation (hierarchy in interface sizes) is conserved throughout evolution. This can be interpreted in at least two different although not mutually exclusive ways: (1) once the complex is formed there is no need to dramatically change the interface size, analogous to the classical explanation for the marginal stability of proteins[20] (that is, selective pressure becomes almost non-existent beyond the point where proteins fold); and (2) maintaining a hierarchy of interface strengths is important for a precise order during assembly[21,22], in which case the largest interface would reflect the main intermediate species during assembly. To test this hypothesis we targeted ten complexes for study using electrospray mass spectrometry (Fig. 4a and Supplementary Table 2).

Initially we verified that the complexes could be generated intact and corresponded to the stoichiometry described in the protein data bank (PDB). The mass spectra recorded for two hexamers with D3 symmetry and one 14-mer with D7 symmetry revealed that the intact homomer is maintained in each case (Fig. 4c). We then induced the disassembly of each complex through the careful change in ionic strength or the stepwise addition of partial denaturants. We detected stable subcomplexes corresponding to trimers and dimers for

hexameric AUH protein (an RNA binding protein), and MoaC (a molybdenum cofactor biosynthesis protein), respectively (Supplementary Table 2). Examination of the interface size shows that in both cases the larger interface is maintained. Similarly for the Ca^{2+}-dependent kinase with D7 symmetry, a dimer is the principal dissociation product and buries the largest interface (Fig. 4c). For one complex (PDB entry 1vea) our results were ambiguous as no intermediate and only monomeric subunits were detected; for another complex (PDB entry 1umg) we predicted a tetramer and detected a dimer. In this case, both subcomplexes bury large surfaces (>5,000Å^2), which may bias the use of interface size as a proxy for interface strength. For the remaining complexes, the predicted subcomplex containing the larger interface was observed. These results demonstrate that the largest interface is maintained consistently during disassembly.

To address whether the disassembly process was the reverse of the assembly pathway, we attempted to reassemble a subset of the complexes studied by dilution of the denaturant and/or manipulation of the ionic strength. In ~50% of the complexes examined we were able to reassemble the original homomer. These results—together with previous studies where reassembly was found to be strongly dependent on factors such as ionic strength, temperature and concentration of denaturant[23,24]—indicate that disassembly is the reverse of assembly under the appropriate conditions.

To complement our experimental observations, we found six additional complexes for which (dis)assembly intermediates had been reported (Fig. 4b). Of these, five match our prediction and one (nucleoside diphosphate kinase) had no intermediate detected. This homomer may either assemble without forming subcomplexes, or subcomplexes may have escaped detection. Alternatively, formation of subcomplexes might involve factors absent from the experimental set-up[25]. Thus, although there are exceptions, we find agreement between the evolutionary pathway and (dis)assembly pathway in 81% of the cases we examined.

Overall, through analysis of a large set of homomers, we have shown that the evolutionary pathway of a homomer can be inferred from its atomic structure morphology. This allowed us to predict the (dis)assembly pathway of homomers in solution, and design mass-spectrometry-based experiments to validate our predictions. Results revealed that the (dis)assembly pathway, which takes place on a protein-folding timescale (~seconds), mimics the evolutionary pathway that has taken place over a considerably longer timescale (~millions of years). This is the first time that a general principle for formation and assembly of homomers has been demonstrated. We hope that this will stimulate further studies, as relationships between

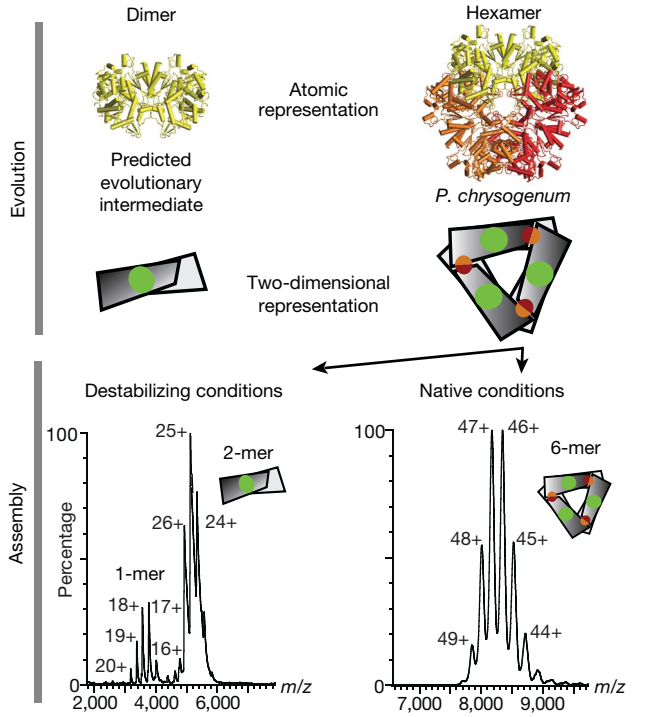


Figure 3 | Prediction of evolutionary routes and link with (dis)assembly in solution. ATP sulphurylase is a hexamer in *Penicillium chrysogenum*, with a predicted dimeric evolutionary intermediate based on interface sizes (that is, the interface in the dimer (green patch) is larger than the trimeric interface (red and orange patch; top panel)). We perturb the hexameric ATP sulphurylase (PDB accession 1m8p, Supplementary Table 2) to disassemble it into subcomplexes probed using electrospray mass spectrometry (bottom panel) and determine whether a dimeric (with the larger interface) or a trimeric (small interface) subcomplex is detected. Both dimers and monomers were detected, with no evidence of trimers.

a

PDB ID	Description	No. of chains	Predicted intermediate	Observed intermediate	Agreement with prediction
1m8p	ATP sulphurylase	6	2	2	✓
1vea	RNA binding AP	6	3	None	-
1hzd	AUH	6,(3)	3	3	✓
1ekr	MoaC	6	2	2	✓
1j2p	proteasome α-ring	7	None	None	✓
1umg	fructose 1,6 biphos.	8	4	2	-
1m3u	KHMase	10	2	2	✓
1pvv	OCTase	12,(9,6)	3	3	✓
1hkx	Ca^{2+}-dependent kinase	14	2	2	✓
1grl	GroEL	14	7	7	✓

b

PDB ID	Description	No. of chains	Predicted intermediate	Observed intermediate	Agreement with prediction
1pfk	PFK I	4	2	2	✓
1nhk	NDK	4	2	None	-
1t3d	serine AT	6	3	3	✓
1aon	GroES	7	None	None	✓
1di0	lumazine synthase	10	5	5	✓
1ogf	RbsD	10	5	5	✓

c

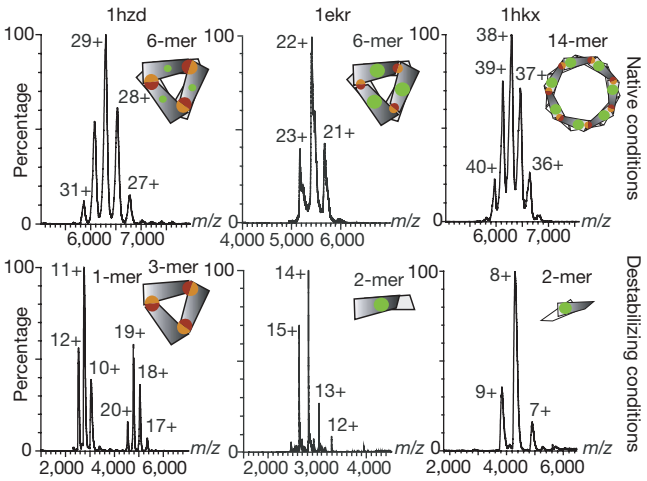


Figure 4 | (Dis)assembly pathways in 16 complexes. a, Homomers for which (dis)assembly was probed using electrospray mass spectrometry (Methods). 'None' denotes no intermediate detected. Complexes agree with our prediction where the subcomplex containing the larger interface is the most stable in eight out of ten cases. **b**, List of homomers for which: (1) information on (dis)assembly has been reported (Supplementary Table 3); (2) a crystal structure is known; and (3) the intermediate species observed during (dis)assembly could be mapped to a subcomplex in the structure. For these complexes, we found an agreement with our prediction in 5/6 cases. **c**, Mass spectra showing intact complexes (top panel) as well as subcomplexes obtained after destabilization in solution (bottom panel). AUH, an RNA binding protein; GroEL and GroES are chaperonins; KHMase, ketopantoate hydroxymethyltransferase; MoaC, molybdenum cofactor biosynthesis protein; RNA binding AP, RNA binding antitermination protein; serine AT, serine acyltransferase; OCTase, ornithine carbamoyltransferase.

folding, complex formation and aggregation are only beginning to be explored.

METHODS SUMMARY

Data set of homomers. All data sets of homomers used were derived from the 3D complex database[8]. As the quaternary structure annotation in the PDB biological unit is erroneous in some cases, we used a manually curated data set[26].

Randomization of evolutionary routes. To assess the significance of the number of evolutionary relationships between proteins with different quaternary structures, we compared the observed numbers to a random model of quaternary structure transitions where evolutionary relationships are reassigned randomly in proportion to the size of each quaternary structure type.

Prediction of evolutionary routes. The size of an interface is given by the number of amino acids in contact, as defined previously[8]. We predict evolutionary intermediates by taking the 'closed' subcomplex containing the largest interface. In cyclic complexes with three or more subunits, each subunit buries two equivalent surfaces. Thus, these interfaces are counted twice when compared to dimer interfaces.

Intact complexes. Complexes were donated by crystallographers and taken from a random selection from the PDB. For further details see Methods.

Generating subcomplexes. Intact complexes were disrupted through change in ionic strength or the stepwise addition of dimethylsulphoxide, methanol or acetonitrile. This process is illustrated in Supplementary Fig. 4 and solution conditions are summarized in Supplementary Table 2.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 15 November 2007; accepted 20 March 2008.
Published online 18 June 2008.

1. Cabezon, E. *et al.* Homologous and heterologous inhibitory effects of ATPase inhibitor proteins on F-ATPases. *J. Biol. Chem.* **277,** 41334–41341 (2002).
2. Hardy, L. W. *et al.* Atomic structure of thymidylate synthase: target for rational drug design. *Science* **235,** 448–455 (1987).
3. Iber, D., Clarkson, J., Yudkin, M. D. & Campbell, I. D. The mechanism of cell differentiation in *Bacillus subtilis*. *Nature* **441,** 371–374 (2006).
4. Marianayagam, N. J., Sunde, M. & Matthews, J. M. The power of two: protein dimerization in biology. *Trends Biochem. Sci.* **29,** 618–625 (2004).
5. Monod, J., Wyman, J. & Changeux, J. P. On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* **12,** 88–118 (1965).
6. Dobson, C. M. Protein folding and misfolding. *Nature* **426,** 884–890 (2003).
7. Hayouka, Z. *et al.* Inhibiting HIV-1 integrase by shifting its oligomerization equilibrium. *Proc. Natl Acad. Sci. USA* **104,** 8316–8321 (2007).
8. Levy, E. D., Pereira-Leal, J. B., Chothia, C. & Teichmann, S. A. 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.* **2,** e155 (2006).
9. Goodsell, D. S. & Olson, A. J. Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 105–153 (2000).
10. Ispolatov, I., Yuryev, A., Mazo, I. & Maslov, S. Binding properties and evolution of homodimers in protein–protein interaction networks. *Nucleic Acids Res.* **33,** 3629–3635 (2005).
11. Pereira-Leal, J. B., Levy, E. D., Kamp, C. & Teichmann, S. A. Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol.* **8,** R51 (2007).
12. Grueninger, D. *et al.* Designed protein–protein association. *Science* **319,** 206–209 (2008).
13. Janin, J. Biochemistry. Dicey assemblies. *Science* **319,** 165–166 (2008).
14. Blundell, T. L. & Srinivasan, N. Symmetry, stability, and dynamics of multidomain and multicomponent protein systems. *Proc. Natl Acad. Sci. USA* **93,** 14243–14248 (1996).
15. Hernandez, H. *et al.* Subunit architecture of multimeric complexes isolated directly from cells. *EMBO Rep.* **7,** 605–610 (2006).
16. Brinda, K. V. & Vishveshwara, S. Oligomeric protein structure networks: insights into protein–protein interactions. *BMC Bioinformatics* **6,** 296 (2005).
17. Monod, J. *Nobel Symposium 11: Symmetry and Function of Biological Systems at the Macromolecular Level* (Almqvist & Wiksell, Stockholm, 1968).
18. Lukatsky, D. B., Shakhnovich, B. E., Mintseris, J. & Shakhnovich, E. I. Structural similarity enhances interaction propensity of proteins. *J. Mol. Biol.* **365,** 1596–1606 (2007).
19. Claverie, P., Hofnung, M. & Monod, J. Sur certaines implications de l'hypothèse d'équivalence stricte entre les protomères des protéines oligomériques. *C. R. Séanc. Acad. Sci.* **266,** 1616–1618 (1968).
20. DePristo, M. A., Weinreich, D. M. & Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Rev. Genet.* **6,** 678–687 (2005).
21. Bahadur, R. P., Rodier, F. & Janin, J. A dissection of the protein-protein interfaces in icosahedral virus capsids. *J. Mol. Biol.* **367,** 574–590 (2007).
22. Powers, E. T. & Powers, D. L. A perspective on mechanisms of protein tetramer formation. *Biophys. J.* **85,** 3587–3599 (2003).
23. Luke, K. & Wittung-Stafshede, P. Folding and assembly pathways of co-chaperonin proteins 10: Origin of bacterial thermostability. *Arch. Biochem. Biophys.* **456,** 8–18 (2006).
24. Cheesman, C., Ruddock, L. W. & Freedman, R. B. The refolding and reassembly of *Escherichia coli* heat-labile enterotoxin B-subunit: analysis of reassembly-competent and reassembly-incompetent unfolded states. *Biochemistry* **43,** 1609–1617 (2004).
25. Kress, W., Mutschler, H. & Weber-Ban, E. Assembly pathway of an AAA+ protein: tracking ClpA and ClpAP complex formation in real time. *Biochemistry* **46,** 6183–6193 (2007).
26. Levy, E. D. PiQSi: Protein quaternary structure investigation. *Structure* **15,** 1364–1367 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the collaborators listed in Supplementary Table 2 for supplying the different complexes and acknowledge H. Hernandez, J. Freeke and L. Lane for assistance with mass spectrometry. We also thank C. Chothia, J. Clark and M. Babu for discussions. This work was supported by the Medical Research Council, the EMBO Young Investigators Programme, the Royal Society and the Waters Kundert Trust.

Author Contributions E.D.L., E.B.E., C.V.R. and S.A.T. designed the experiments and wrote the manuscript; E.D.L. and E.B.E. performed the bioinformatics and mass spectrometry experiments, respectively.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to E.D.L., C.V.R. or S.A.T. (homomers@mrc-lmb.cam.ac.uk).

METHODS

Data set of homomers and symmetry information. The ~5,000 structures data set used throughout the study is non-redundant at 80% sequence identity. The data set was controlled for a possible bias in the distribution of the number of subunits. As no bias was found[8], we can be confident in the accuracy of the relative abundances of symmetries as well as their evolutionary relationships. However, an important bias in structural data is the under-representation of membrane proteins, which is discussed further in Supplementary Information 2. For the analysis on quaternary structure conservation, we derived several non-redundant sets of protein pairs. To study conservation within the identity range $X\% - X + 10\%$, we used a data set non-redundant at $X + 20\%$ (with the exception of $X = 90$ and 100%). All data sets and the symmetry information were derived from the 3D complex database[8].

Randomization of evolutionary routes. To assess the significance of the number of evolutionary relationships between proteins with different quaternary structures, we devised a random model of quaternary structure transitions. In this model, evolutionary relationships are reassigned randomly in proportion to the size of each quaternary structure type. For each evolutionary link between two quaternary structure types, a first quaternary structure type is picked up with a probability $p(\mathrm{QS}) = T^{\mathrm{QS}} / T$, where T^{QS} is the quaternary structure size (number of proteins), and T is the total number of proteins. A second quaternary structure is chosen in the same way but the type picked first is set aside and cannot be selected again. One-hundred rounds of reassignment were performed, and a mean number of links and associated standard deviation were calculated for each quaternary structure pair.

Prediction of evolutionary routes. To decompose the complexes into their evolutionary intermediates, we first grouped together interfaces related by a symmetry operation. We then ranked each group according to the average size of interfaces it contained. The size is given by the number of amino acids in contact as defined previously[8]. The complex was broken by removing each group of interfaces one by one, starting with the weakest. After removal of each group, we checked if all the subunits in the complex were still connected via the remaining interfaces. When the complex breaks down, the subcomplexes found correspond to the predicted evolutionary intermediates. Note that in cyclic complexes with three subunits or more, each subunit buries two equivalent surfaces. Thus, these interfaces are counted twice when compared to interfaces within dimers

Probing the (dis)assembly pathway using electrospray mass spectrometry. Methanol and acetonitrile were obtained from Fisher scientific; ammonium acetate and dimethylsulphoxide were from Sigma. All chemicals used were American Chemical Society or HPLC grade and water was obtained from an ELGA LabWater's PURELAB Maxima system.

Before mass spectrometry, complex-containing solutions were desalted and concentrated by centrifugation at 10,000*g* in Vivaspin concentrator tubes (exclusion limits 5,000, 10,000, 30,000; Vivaspin, Sartorius) to a final concentration of 20–55 μM of protein complex (Supplementary Table 2). The intact complex was diluted with ammonium acetate to 3–5 μM immediately before the mass spectrometry analysis. Two microlitres of complex-containing solutions were analysed using nanoelectrospray and a quadrupole time-of-flight mass spectrometer (QSTAR, Sciex). The instrument was modified for the detection of high masses[27]. For nanoelectrospray, gold-coated borosilicate capillaries were prepared in-house as described previously[28]. The following instrumental parameters were used: capillary voltage up to 1.5 kV, declustering potential 200 V, focusing potential 250 V, declustering potential-2 15 V and collision energy up to 280 V, microchannel plate detector 2350V. Argon was used as a collision gas for tandem mass spectra. All spectra were calibrated externally using caesium iodide (100 mg ml^{-1}).

27. Sobott, F. *et al.* A tandem mass spectrometer for improved transmission and analysis of large macromolecular assemblies. *Anal. Chem.* **74,** 1402–1407 (2002).
28. Hernandez, H. & Robinson, C. V. Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. *Nature Protocols* **2,** 715–726 (2007).

LETTERS

Modest stabilization by most hydrogen-bonded side-chain interactions in membrane proteins

Nathan HyunJoong Joh[1], Andrew Min[1], Salem Faham[2], Julian P. Whitelegge[3], Duan Yang[1], Virgil L. Woods Jr[4] & James U. Bowie[1]

Understanding the energetics of molecular interactions is fundamental to all of the central quests of structural biology including structure prediction and design, mapping evolutionary pathways, learning how mutations cause disease, drug design, and relating structure to function. Hydrogen-bonding is widely regarded as an important force in a membrane environment because of the low dielectric constant of membranes and a lack of competition from water[1–6]. Indeed, polar residue substitutions are the most common disease-causing mutations in membrane proteins[6,7]. Because of limited structural information and technical challenges, however, there have been few quantitative tests of hydrogen-bond strength in the context of large membrane proteins. Here we show, by using a double-mutant cycle analysis, that the average contribution of eight interhelical side-chain hydrogen-bonding interactions throughout bacteriorhodopsin is only 0.6 kcal mol^{-1}. In agreement with these experiments, we find that 4% of polar atoms in the non-polar core regions of membrane proteins have no hydrogen-bond partner and the lengths of buried hydrogen bonds in soluble proteins and membrane protein transmembrane regions are statistically identical. Our results indicate that most hydrogen-bond interactions in membrane proteins are only modestly stabilizing. Weak hydrogen-bonding should be reflected in considerations of membrane protein folding, dynamics, design, evolution and function.

The few evaluations of hydrogen-bond contributions in membrane proteins have tested the effect of single point mutants on either the free energy of unfolding or the free energy of dissociation[4,8,9]. However, these measurements combine hydrogen-bond contributions with desolvation and many other factors[10], so the hydrogen-bond contribution cannot necessarily be extracted without the incorporation of correction factors[11] that are particularly uncertain for membrane proteins.

The energetic complexities of single side-chain alterations can be illustrated by mutations in bacteriorhodopsin residues T90 and D115 that make two hydrogen bonds near the centre of the membrane (Fig. 1). We eliminated the hydrogen bonds by making T90A and D115A mutations and measured the change in the free energy of unfolding with an SDS unfolding assay[9]. The T90A mutation decreases stability by 1.3 ± 0.1 kcal mol^{-1}, whereas the D115A mutant increases stability by 0.5 ± 0.1 kcal mol^{-1}. The large variation suggests that hydrogen-bonding alone does not dominate the stability effects, and other energetic contributions must be accounted for. Below we present evidence that a principal factor is changes in solvation free energy in the unfolded protein.

To examine the effects of the T90A and D115A mutations on the folded state of bacteriorhodopsin, we solved the structures of the D115A mutant and a T90A/D115A double mutant (T90A proved too unstable to crystallize). We were unable to detect any structural changes in the mutant proteins that would obviously explain the contrasting energetic consequences, beyond the loss of density around the deleted side chains (see Fig. 2a).

To probe the consequences of the mutations on the unfolded state, we developed a hydrogen-exchange assay. Unfolded-state backbone hydrogens that are shielded from solvent by burial in the detergent micelle will exchange at a slower rate than backbone hydrogens exposed to the aqueous phase[12,13]. Figure 2b shows the detailed time course of exchange for the unfolded state of the wild-type and mutant proteins at three regions, one resolved by the peptide overlapping the site of the T90A mutation, the second overlapping a region in between the sites of the T90A and D115A mutations, and the third overlapping the site of the D115A mutation. Figure 2c summarizes the average exchange rates of peptides throughout the unfolded states.

The T90A mutation modestly slows the exchange in the vicinity of position 90, whereas D115A markedly slows exchange in the vicinity of position 115. Although the sequence effects on intrinsic exchange rates[14] are uncertain in an SDS environment[15], the results suggest that the polar to non-polar substitutions alter the unfolded state by increasing burial in the detergent micelle at the sites of mutation. The larger change in polarity in D115A than in T90A is consistent with the larger effect on exchange rate and probably explains the stabilizing effect of the D115A mutation. In particular, the loss of the favourable escape of D115 to solvent could increase the free energy of the unfolded state in the D115A mutant, compensating for the increased free energy of the folded state. Thus, solvation effects in the unfolded state may mask the hydrogen-bond contribution that we wish to measure.

In an effort to obtain side-chain interaction energies within the folded state, we turned to double-mutant cycle analysis. Double-mutant cycle analysis has the potential to measure the free energy of side-chain interaction directly in the context of the folded protein by cancelling out energetic perturbations in both the folded and unfolded states that are not due to the interactions between the side chains[16,17]. Thus, desolvation contributions and any other new interactions made in the unfolded state can be eliminated. As a result, double-mutant cycle analysis can be interpreted as reporting the contribution of the hydrogen-bonded interaction to the free energy of the folded state, not the difference in free energy between the folded and unfolded states. The unfolded state becomes simply a common reference state in which the interaction of interest is broken (see Supplementary Methods).

We were able to express and purify complete single-mutant and double-mutant sets for eight interhelical hydrogen-bonding

[1]Department of Chemistry and Biochemistry, UCLA-DOE Center for Genomics and Proteomics, Molecular Biology Institute, [2]Department of Physiology, and [3]The NPI-Semel Institute, Pasarow Mass Spec Laboratory, University of California, Los Angeles, California 90095, USA. [4]Department of Medicine and Biomedical Sciences Graduate Program, University of California, San Diego, La Jolla, California 92093-0656, USA.

interactions as shown in Fig. 1. Four of the hydrogen bonds are in the middle of the hydrocarbon core region of the bilayer, three are on the edge of the hydrocarbon core and one is in the interfacial region. The strongest interactions were T46–D96 and T90–D115, each contributing -1.7 ± 0.3 kcal mol^{-1}, and T170–S226, contributing -0.8 ± 0.3 kcal mol^{-1}. The strongest interactions, between T46 and D96 and between T90 and D115, both involve two hydrogen bonds, corresponding to about -0.9 kcal mol^{-1} per hydrogen bond. Y185–D212 and S193–E204 make weaker, but favourable, interactions contributing -0.4 ± 0.4 and -0.5 ± 0.3 kcal mol^{-1}, respectively. The K30–Y43 and E9–Y79 interactions were found to make no measurable contribution to stability, and W189–Y83 was found to be slightly destabilizing, contributing $+0.4 \pm 0.2$ kcal mol^{-1}.

The results of the double-mutant cycles suggest three main conclusions. First, hydrogen-bonded side-chain contributions are quite variable and depend on the characteristics and local environment of each hydrogen bond. Second, the strength of a hydrogen-bonded interaction is not strongly correlated with the location in the protein. For example, the T170–S226 interaction in the interfacial region contributes -0.8 ± 0.3 kcal mol^{-1}, whereas the T185–D212 interaction in the centre of the hydrocarbon core contributes only -0.4 ± 0.4 kcal mol^{-1}. Third, the eight hydrogen-bonding interactions studied here make a remarkably modest average contribution of only about 0.6 kcal mol^{-1}, which corresponds to a roughly threefold effect on an equilibrium constant at room temperature.

Protein folding experiments are complex and not all variables can be eliminated, so we sought an additional, independent evaluation of the hydrogen-bond contribution in membrane proteins. We reasoned that if hydrogen-bond strengths were low, we would see a large number of unsatisfied hydrogen-bonding groups in membrane protein structures[18]. To test this idea, we examined six membrane protein structures solved at 1.7 Å resolution or better. HBPLUS was used to identify the hydrogen-bonding of all polar atoms within the hydrocarbon core region of the bilayer. Any polar atoms that made no hydrogen bonds were further verified by eye.

The results are summarized in Table 1 and reveal that unsatisfied hydrogen bonds are not rare in the hydrocarbon core region (also see Supplementary Information). Of 2,892 protein donors and acceptors examined in the hydrocarbon core region, 111 have no hydrogen-bonding partner (about 4%). We believe this to be a low estimate of the number of unsatisfied hydrogen bonds because the HBPLUS criteria permit even marginal hydrogen bonds to be counted. Moreover, the crystal structure reports the predominant conformation and does not report the fraction of time for which a hydrogen bond is broken.

The hydrogen-bonded interaction strengths we measured in a membrane protein are very similar to hydrogen-bond strengths in

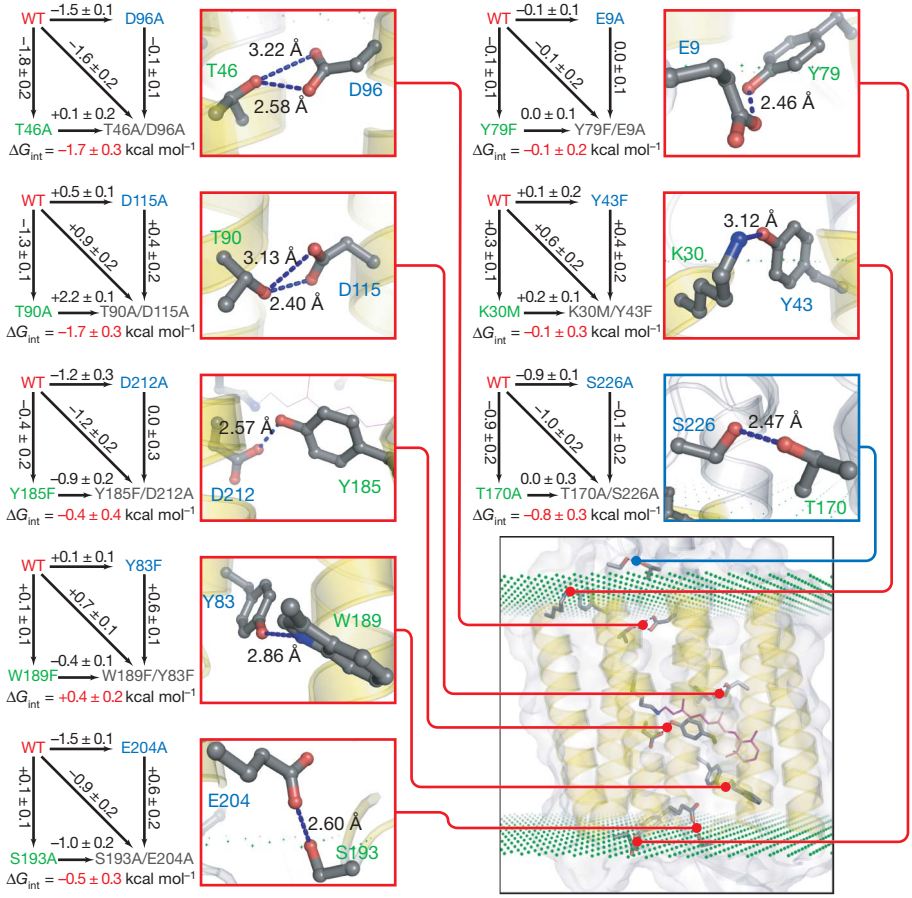


Figure 1 | Double-mutant cycles for hydrogen-bonding interactions in bacteriorhodopsin. For each cycle shown, the difference in free energies of unfolding (black number by the arrow) was measured for the pair of proteins connected by the arrow. Free energies of unfolding are compared at an SDS concentration at which the wild-type protein (WT) is 50% unfolded to minimize extrapolations needed. Errors are s.d. for three separate measurements. Next to each double-mutant cycle is a close-up view of the relevant hydrogen bond shown as blue dotted line between the altered side chains along with the heavy atom donor–acceptor distance. Donor and acceptor residues are labelled in green and blue, respectively. Donor–acceptor distinction in the two strongest interactions was arbitrary. On the basis of hydrogen-bonding patterns and nearest neighbours, it seems that all the potentially charged residues are the neutral species. The inset (bottom right) shows the location of each interaction in the context of the protein (PDB ID 1C3W). The planes of green dots indicate the estimated position of the edge of the hydrocarbon region of the bilayer as defined previously[28]. Any interaction mediated by the residues that contain at least one atom in the hydrocarbon region is mapped with the red line, and the interaction in the lipid/water interface region is mapped with a blue line.

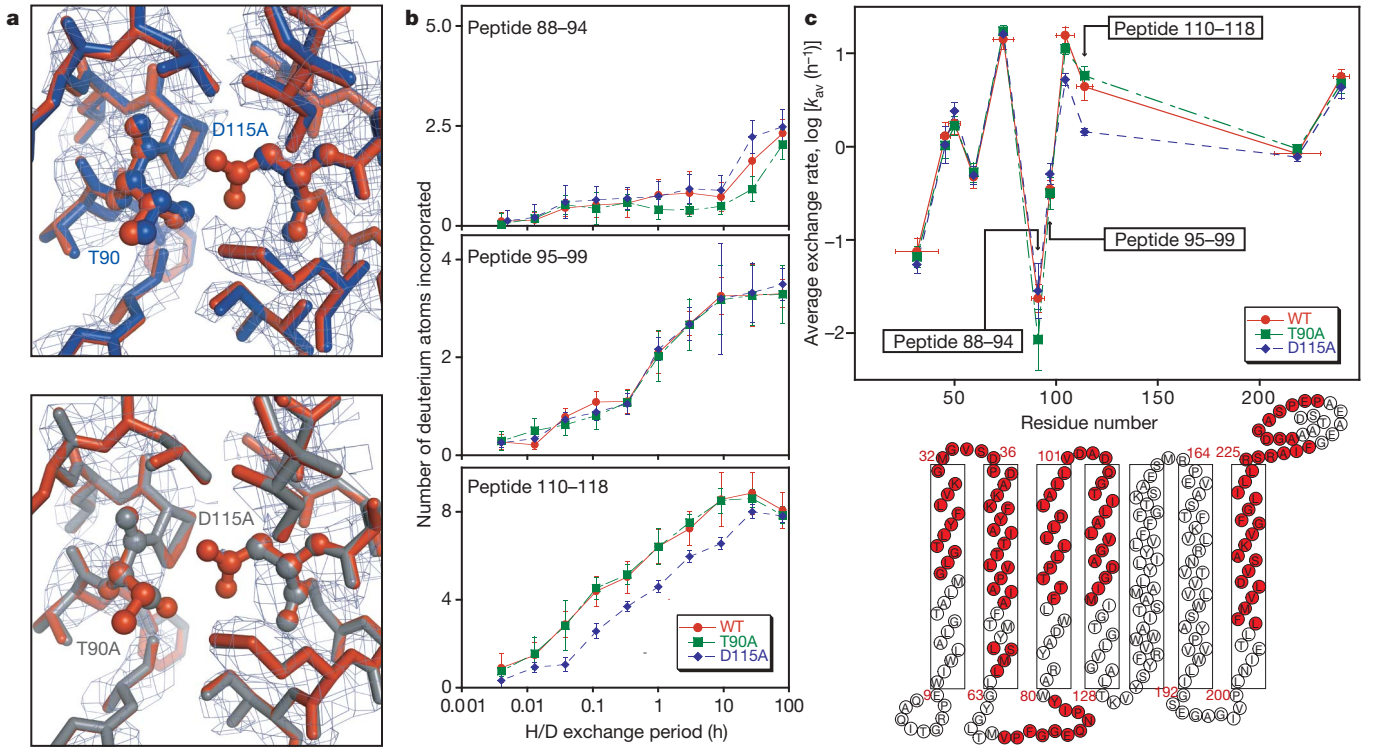


Figure 2 | Characterization of the T90A, D115A and T90A/D115A mutants. **a**, Omit electron density maps and overlay of refined mutant and wild-type structures for D115A (top) and T90A/D115A (bottom) mutants. The wild-type structure (PDB ID 1PY6) is shown in rust, D115A in blue and T90A/D115A in grey. The mutated side chains are shown in ball-and-stick representation and labelled. The side chains of all residues within 4 Å of T90 and D115 of the wild-type (WT) structure were eliminated during refinement for the omit map and are shown here with the exception of W182, which was left out for clarity. The electron density map is contoured at 1.0σ and 1.5σ for D115A and T90A/D115A, respectively. **b**, Plot of the number of hydrogens exchanged in the denatured state against time for peptides overlapping the T90A mutation (top), a region between T90A and D115A mutation (middle) and the D115A mutation (bottom). In brief, wild-type and mutant proteins were unfolded in SDS and incubated in D_2O; the exchange reaction was quenched by rapidly lowering the temperature and pH. The proteins were then digested with pepsin, distinct peptides were separated chromatographically and the change in the mass envelope was measured by electrospray ionization-mass spectroscopy. The maximum scale on the *y* axis is the maximum number of exchangeable backbone amide hydrogens. Error bars are s.d. estimated with results from triplicate experiments. **c**, A plot of average exchange rates for peptides throughout the protein (top) and a schematic illustration of the bacteriorhodopsin structure (bottom) showing the sequences covered by the deuterium exchange experiment in light red. Error bars on the *x* axis reflect the range of the peptic peptides, and those on the *y* axis are s.d. for ten simulated data sets incorporating the experimental errors observed in the exchange time courses (see Methods).

soluble proteins measured in a variety of double-mutant cycle analyses (see Supplementary Fig. 1). Retrospectively, this finding is not unreasonable because the polarities of the interiors of soluble proteins and membrane proteins are quite similar[19,20]. Because folding studies in soluble proteins are well accepted, we decided to validate our findings further by comparing hydrogen-acceptor distance distributions in membrane and soluble proteins. As summarized in Fig. 3, the buried hydrogen-bond distances in the interior of soluble and membrane protein transmembrane regions are statistically indistinguishable (see Supplementary Fig. 3 for full distributions), both averaging 2.02 Å. However, the hydrogen-bond distances in surface residues are markedly different. For the transmembrane regions of membrane proteins, the hydrogen-acceptor distances on the surface are slightly shortened to 1.98 Å on average, whereas for soluble proteins the average distance lengthens to 2.08 Å. These results further validate our results, indicating similar contributions from interior hydrogen bonds in soluble and membrane proteins. It also hints that hydrogen bonds at the surface of membrane proteins may be stronger than what we have measured here for interior ones.

Many of the hydrogen-bonded residues we tested are involved in function, so it is possible that they are in a separate class from structural hydrogen-bonded side chains. However, previous work eliminating hydrogen bonds in structural residues is consistent with our findings. For example, a Gln residue that makes two hydrogen bonds across the interface of an OMPLA dimer contributes less than 1 kcal mol^{-1} to dimerization (less than 0.5 kcal mol^{-1} per hydrogen bond)[21]. An Asp to Ala substitution in a designed transmembrane helix oligomer decreases the free energy of association by 1.8 kcal mol^{-1} (0.9 kcal mol^{-1} per potential hydrogen bond)[4]. A T87A substitution in glycophorin A decreases the free energy of

Table 1 | Unsatisfied hydrogen-bond donors and acceptors in hydrophobic cores of membrane proteins

Protein	PDB code	Resolution (Å)	Donor and acceptor population, unsatisfied/total*	Percentage unsatisfied
Bacteriorhodopsin	1C3W	1.55	17/345	4.9
Formate dehydrogenase N	1KQF	1.60	9/288	3.1
NH_4^+ transporter Amt B	1U7G	1.40	17/571	3.0
Na^+/Cl^--dependent neurotransporter transmitter	2A65	1.65	34/699	4.9
NH_4^+ transporter Amt-1	2B2H	1.54	24/593	4.0
Aquaporin	2F2B	1.68	10/396	2.5
Total			111/2,892	3.8

* See Supplementary Fig. 2 for further details.

dimerization by 0.9 kcal mol^{-1} (about 0.5 kcal mol^{-1} per hydrogen bond in the dimer)[22]. Mutations in hydrogen-bonding polar residues in the T-cell receptor ζ-subunit dimer affected dimerization by a maximum of 7.7-fold, as indirectly measured by assembly rate, which corresponds to a maximum of about 1.2 kcal mol^{-1} (0.6 kcal mol^{-1} per residue in the dimer)[8]. Analysis of hydrogen-bonding mutations in bacteriorhodopsin (ref. 9) suggest a contribution of about 1 kcal mol^{-1} (ref. 23). Because double-mutant cycle analysis was not employed in these cases, however, the results combine hydrogen-bond contributions with other effects that could contribute favourably or unfavourably[10,24].

Although our results indicate that most hydrogen-bond interactions observed in membrane proteins make modest energetic contributions, it does not mean that polar interactions cannot be strong. It has been found[16] that charge-stabilized salt-bridge interactions can contribute 5.6 kcal mol^{-1}, and mutations in residues that hydrogen-bond to ligands in the β-adrenergic receptor can have marked effects on ligand binding[25].

Why, then, are hydrogen-bond interactions not much stronger on average? It is possible that optimal geometries are difficult to achieve, that there are entropic costs to fixing hydrogen-bonded groups, and that polar groups in the protein can increase the local dielectric constant[26]. In addition to possible physical limitations, there may also be evolutionary pressure favouring weak hydrogen bonds. Evolutionary pressure for weak hydrogen bonds could come in the form of the conformational flexibility needed for protein function[6]. Moreover, the helical distortions that are common in membrane proteins would be hard to create by random mutation if the breakage of hydrogen bonds presented a large energy barrier. It is also possible that weak hydrogen bonds are more robust evolutionarily. Strong side-chain hydrogen bonds, once established, could no longer be altered by mutation without destroying fitness. Thus, proteins that rely on a strong hydrogen bond for stability would be more likely to be lost from a population than proteins that rely on more broadly distributed stabilizing interactions. Whatever the mechanism, our results indicate that it is not difficult to make and break interactions between polar residues, enabling the structural variation and dynamic flexibility necessary to optimize membrane protein function and folding. The results also suggest that a primary mechanism for the prevalence of disease-causing substitutions of polar residues may not be inappropriate hydrogen-bond formation but, instead, alterations in bilayer partitioning[6]. However, the loss or gain of even a weak hydrogen bond could tip the balance between biological function and dysfunction.

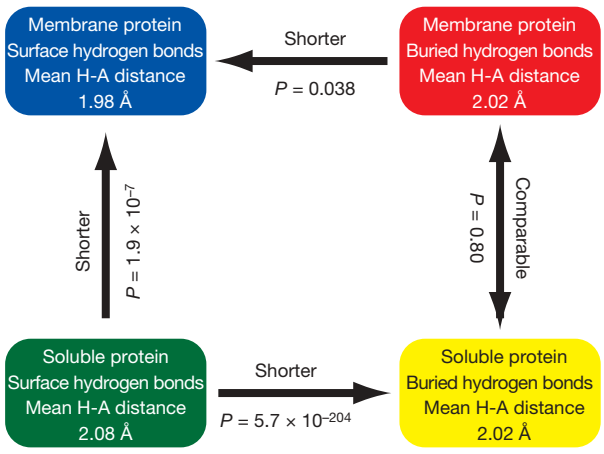


Figure 3 | Comparison of average hydrogen-bond distances in different environments. The arrows point towards the shorter hydrogen bonds. The *P* value is the probability that the distance distributions are different by random chance based on Student's *t*-test. The distributions are shown in Supplementary Information.

METHODS SUMMARY

Equilibrium unfolding measurements. Stability measurements were performed essentially as described previously[9]. In brief, purple membrane was dissolved in a DMPC (1,2-dimyristoyl-*sn*-glycerol-3-phosphocholine)/CHAPSO (3((3-cholamidopropyl)dimethylammonio)-2-hydroxy-1-propanesulphonate) mixture and unfolded by adding increasing concentrations of SDS. Unfolding was monitored either by retinal absorbance at 560 nm or far-ultraviolet circular dichroism at 228 nm. Unfolding free energies were compared at the SDS concentration at which wild-type bacteriorhodopsin was 50% unfolded, to minimize the extrapolation error due to the varying *m* values.

X-ray crystallography. Crystals were grown by the bicelle method[27] and, diffraction data were phased by molecular replacement.

Deuterium exchange. Unfolded proteins in SDS were deuterium-exchanged for various periods. The exchange reactions were then quenched by rapid cooling and the addition of low-pH buffer containing an acid-labile detergent to maintain solubility during digestion with pepsin. K^+ was also included to precipitate SDS. Quenched reactions were flash-frozen and stored at −80 °C. For analysis of deuterium exchange, the quenched reactions were rapidly thawed, treated with pepsin at 0 °C and immediately analysed by liquid chromatography–mass spectrometry (LC–MS).

Structure analysis. To analyse unsatisfied hydrogen-bond donors and acceptors, the transmembrane regions were first identified as described[28] and hydrogen bonds were identified with HBPLUS[29]. To obtain hydrogen-bond distance distributions, hydrogen-acceptor distances of all α-helix backbone–backbone hydrogen bonds in six high-resolution membrane protein structures and 839 unique soluble protein structures were calculated with HBPLUS. The soluble proteins chosen were solved in the same resolution range as the membrane proteins used.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 24 January; accepted 8 April 2008.
Published online 25 May 2008.

1. White, S. H. How hydrogen bonds shape membrane protein structure. *Adv. Protein Chem.* **72,** 157–172 (2005).
2. Popot, J. L. & Engelman, D. M. Helical membrane protein folding, stability, and evolution. *Annu. Rev. Biochem.* **69,** 881–922 (2000).
3. Zhou, F. X., Merianos, H. J., Brunger, A. T. & Engelman, D. M. Polar residues drive association of polyleucine transmembrane helices. *Proc. Natl Acad. Sci. USA* **98,** 2250–2255 (2001).
4. Gratkowski, H., Lear, J. D. & DeGrado, W. F. Polar side chains drive the association of model transmembrane peptides. *Proc. Natl Acad. Sci. USA* **98,** 880–885 (2001).
5. Adamian, L. & Liang, J. Interhelical hydrogen bonds and spatial motifs in membrane proteins: polar clamps and serine zippers. *Proteins* **47,** 209–218 (2002).
6. Partridge, A. W., Therien, A. G. & Deber, C. M. Polar mutations in membrane proteins as a biophysical basis for disease. *Biopolymers* **66,** 350–358 (2002).
7. Partridge, A. W., Therien, A. G. & Deber, C. M. Missense mutations in transmembrane domains of proteins: phenotypic propensity of polar residues for human disease. *Proteins* **54,** 648–656 (2004).
8. Call, M. E. *et al.* The structure of the ζζ transmembrane dimer reveals features essential for its assembly with the T cell receptor. *Cell* **127,** 355–368 (2006).
9. Faham, S. *et al.* Side-chain contributions to membrane protein structure and stability. *J. Mol. Biol.* **335,** 297–305 (2004).
10. Duong, M. T., Jaszewski, T. M., Fleming, K. G. & MacKenzie, K. R. Changes in apparent free energy of helix–helix dimerization in a biological membrane due to point mutations. *J. Mol. Biol.* **371,** 422–434 (2007).
11. Myers, J. K. & Pace, C. N. Hydrogen bonding stabilizes globular proteins. *Biophys. J.* **71,** 2033–2039 (1996).
12. Busenlehner, L. S. & Armstrong, R. N. Insights into enzyme structure and dynamics elucidated by amide H/D exchange mass spectrometry. *Arch. Biochem. Biophys.* **433,** 34–46 (2005).
13. Busenlehner, L. S. *et al.* Stress sensor triggers conformational response of the integral membrane protein microsomal glutathione transferase 1. *Biochemistry* **43,** 11145–11152 (2004).
14. Molday, R. S., Englander, S. W. & Kallen, R. G. Primary structure effects on peptide group hydrogen exchange. *Biochemistry* **11,** 150–158 (1972).
15. O'Neil, J. D. & Sykes, B. D. NMR studies of the influence of dodecyl sulfate on the amide hydrogen exchange kinetics of a micelle-solubilized hydrophobic tripeptide. *Biochemistry* **28,** 699–707 (1989).
16. Hong, H., Szabo, G. & Tamm, L. K. Electrostatic couplings in OmpA ion-channel gating suggest a mechanism for pore opening. *Nature Chem. Biol.* **2,** 627–635 (2006).
17. Fersht, A. R., Matouschek, A. & Serrano, L. The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* **224,** 771–782 (1992).
18. Fleming, P. J. & Rose, G. D. Do all backbone polar groups in proteins form hydrogen bonds? *Protein Sci.* **14,** 1911–1917 (2005).
19. Adamian, L., Nanda, V., DeGrado, W. F. & Liang, J. Empirical lipid propensities of amino acid residues in multispan alpha helical membrane proteins. *Proteins* **59,** 496–509 (2005).

20. Rees, D., DeAntonio, L. & Eisenberg, D. Hydrophobic organization of membrane proteins. *Science* **245,** 510–513 (1989).
21. Stanley, A. M. & Fleming, K. G. The role of a hydrogen bonding network in the transmembrane β-barrel OMPLA. *J. Mol. Biol.* **370,** 912–924 (2007).
22. Fleming, K. G. & Engelman, D. M. Specificity in transmembrane helix–helix interactions can define a hierarchy of stability for sequence variants. *Proc. Natl Acad. Sci. USA* **98,** 14340–14344 (2001).
23. Senes, A., Engel, D. E. & DeGrado, W. F. Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. *Curr. Opin. Struct. Biol.* **14,** 465–479 (2004).
24. Lear, J. D., Gratkowski, H., Adamian, L., Liang, J. & DeGrado, W. F. Position-dependence of stabilizing polar interactions of asparagine in transmembrane helical bundles. *Biochemistry* **42,** 6400–6407 (2003).
25. Rosenbaum, D. M. *et al.* GPCR engineering yields high-resolution structural insights into β_2-adrenergic receptor function. *Science* **318,** 1266–1273 (2007).
26. Shan, S. O. & Herschlag, D. Hydrogen bonding in enzymatic catalysis: analysis of energetic contributions. *Methods Enzymol.* **308,** 246–276 (1999).
27. Faham, S. & Bowie, J. U. Bicelle crystallization: a new method for crystallizing membrane proteins yields a monomeric bacteriorhodopsin structure. *J. Mol. Biol.* **316,** 1–6 (2002).
28. Chamberlain, A. K., Lee, Y., Kim, S. & Bowie, J. U. Snorkeling preferences foster an amino acid composition bias in transmembrane helices. *J. Mol. Biol.* **339,** 471–479 (2004).
29. McDonald, I. K. & Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238,** 777–793 (1994).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the staff at the beamlines 8.2.1 and 8.2.2 at the Advanced Light Source; F. Pettit for advice on statistics; M. Philips for assisting with circular dichroism experiments; S. Bassilian for assisting with LC–MS analysis of intact bacteriorhodopsin; Y. Ihm for the identification of transmembrane regions; Z. Zhang for providing deuterium-exchange data reduction software MAGTRAN and LAPLACE; N. L. Kelleher for providing the acid-labile detergent; and M. Chamberlain, H. Cheng, E. Gendel, H. Hong, Y. Ihm, A. D. Meruelo, T. Mitchell and R. Stafford for critically reading the manuscript. This work was supported by National Institutes of Health grant RO1 GM063919 (J.U.B.) and by the National Institutes of Health National Cancer Institute Innovative Molecular Analysis Technologies Program (V.L.W.).

Author Contributions N.H.J. and J.U.B. designed the research and prepared the manuscript. N.H.J. performed the vast majority of the experiments and structure analyses. A.M. and D.Y. assisted with mutagenesis and protein purification. A.M. crystallized the T90A/D115A mutant. S.F. collected and processed some diffraction data and helped with structure determination and refinement. J.P.W. assisted site-directed mutagenesis verification by LC–MS analysis of intact bacteriorhodopsin, provided technical advice on mass spectrometry, and helped develop the H/D exchange method. V.L.W. assisted in H/D exchange data analysis, including the provision of specialized software and hardware, and provided help with H/D exchange methods.

Author Information Coordinates and structure factors for the D115A and T90A/D115A mutant bacteriorhodopsins have been deposited in the Protein Data Bank under accession codes 3COC and 3COD, respectively. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.U.B. (bowie@mbi.ucla.edu).

METHODS

SDS unfolding assays. All bacteriorhodopsin mutants were created and purified as described[9]. The SDS unfolding assays were performed as described previously[9] for all except two of the mutant proteins by titration in 20% (w/v) SDS in 15 mM DMPC, 6 mM CHAPSO and 10 mM sodium phosphate pH 6.0 into 0.1 mg ml^{-1} protein suspended in 15 mM DMPC, 6 mM CHAPSO and 10 mM sodium phosphate pH 6.0. For the D212A and Y185F/D212A mutants, retinal absorbance was unusually sensitive to SDS concentration, causing a large slope to the native baseline of the unfolding curves that made interpretation difficult. We therefore turned to far-ultraviolet circular dichroism. Circular dichroism was monitored at 228 nm to minimize absorbance. The mutant bacteriorhodopsins D212A and Y185F/D212A were unfolded in a 1-cm path length quartz cuvette by additions of 10% (w/v) SDS in 7.5 mM DMPC, 3 mM CHAPSO and 10 mM sodium phosphate pH 6.0 into 0.1 mg ml^{-1} protein prepared in 7.5 mM DMPC, 3 mM CHAPSO and 10 mM sodium phosphate pH 6.0.

X-ray crystallography. Mutant bacteriorhodopsins D115A and T90A/D115A were crystallized by using the bicelle method[27]. A 4:1 protein/bicelle solution (4 µl) of 10 mg ml^{-1} purple membrane in water and 40% (w/v) 2.8:1 DMPC/CHAPSO was mixed with 1.5 µl of a well solution, inverted over the well solution, and incubated at 37 °C. D115A was crystallized by using a well solution containing 1.4 M NaH_2PO_4 pH 3.7, 0.8 M NaH_2PO_4 pH 4.5 and 0.12 M hexanediol. The well solution for T90A/D115A was 2.4 M NaH_2PO_4 pH 3.7, 3.5% triethylene glycerol and 0.15 M hexanediol. Crystals were transferred to a cryoprotectant solution of 4.0 M NaH_2PO_4 pH 3.7 before being frozen under a stream of liquid nitrogen. Data were collected from D115A and T90A/D115A crystals at the Advanced Light Source beamlines 8.2.2 and 8.2.1, respectively, at the Lawrence–Berkeley National Laboratory. X-ray diffraction data were processed and scaled by using the DENZO package[30], and the structure was solved by molecular replacement with the use of the 1PY6 structure[9] and refined by using CNS with a twinning fraction of 0.50 (ref. 31) at resolutions of 2.3 Å and 2.7 Å for D115A and T90A/D115A, respectively. Five per cent of the reflections were withheld for the calculation of R_{free}. The same reflections withheld in the WT refinement were used for both mutant refinements to eliminate bias. D115A was refined to an R factor of 22.1 and an R_{free} of 27.5, and T90A/D115A was refined to an R factor of 27.3 and an R_{free} of 28.7. See Supplementary Table 2 for detailed data collection and refinement statistics.

Hydrogen–deuterium exchange analysis of the unfolded state. To initiate the deuterium exchange reaction, nine parts of deuterium exchange (DX) buffer (10 mM sodium phosphate in D_2O (uncorrected pH 6.0)) were added to 1 part of SDS-denatured protein: 5 mg ml^{-1} in 5.5% (w/v) SDS, 15 mM DMPC, 6 mM CHAPSO and 10 mM sodium phosphate pH 6.0. The exchange reaction was incubated at 22 °C and quenched at various time points by transferring a 40-µl aliquot of the exchange reaction into a chilled Microfuge tube at 0 °C containing a 10-µl droplet of 1.0 M potassium phosphate pH 2.4 and 2.5% (w/v) acid-labile surfactant, followed by rapid mixing and immediate flash freezing in liquid N_2. The K^+ precipitates dodecylsulphate, which would otherwise inhibit the subsequent pepsinization. Acid-labile surfactant helps to solubilize bacteriorhodopsin and peptic peptides, and is degraded during subsequent low-pH chromatography. The quenched reaction was stored at −80 °C until LC–MS analyses. No significant change in the deuterium content in protein takes place at −80 °C (data not shown).

For proteolysis, the flash-frozen quenched exchange reaction was rapidly thawed to 0 °C and immediately transferred to a spin filter (SpinX, 0.22-µm filter; Costar) chilled to 0 °C and containing 75 µl of immobilized pepsin slurry (Pierce) prewashed in DX buffer and dried by low-speed centrifugation. After an incubation for 9 min at 0 °C, the samples were immediately centrifuged at 10,000*g* at 0 °C for 30 s to collect the peptic peptides in the filtrate. The filtrate was then immediately injected into an LC–MS column with a chilled syringe.

The peptide mixture was applied to a PLRP-S 100-Å column (1 mm × 50mm; Polymer Laboratory) equilibrated in 95% solvent A (0.4% formic acid, 0.1% hexafluoroisopropanol (HFIP), 99.5% water) and 5% solvent B (0.4% formic acid, 0.1% HFIP, 99.5% acetonitrile). After washing in the same solvent mixture for 2.15 min at a flow rate of 120 µl min^{-1}, the flow rate was changed to 60 µl min^{-1} and a gradient initiated to 95% solvent B over 22.85 min. The column was cleaned at the end of each experiment by injecting 50 µl of formic acid while eluting with 95% solvent B at 120 µl min^{-1} for 2 min, followed by equilibration with 5% solvent B at 120 µl min^{-1} for 5 min. To minimize back exchange, LC–MS was performed at pH 2.4 and all LC–MS components starting from the outlet from pump to the inlet of ion source were completely immersed in a tightly packed ice bath. ThermoFinnigan LCQ was used to collect the isotopic envelope evolution data.

The average number of amide hydrogen atoms exchanged for deuterium was determined from the measured shift in the centroid mass for each isotopic distribution. To account for back exchange and forward exchange during analysis, the deuterium content, D, in each fragment was calculated from

$$D = N(m - m_{0\,\mathrm{min}})/(m_{\mathrm{equilibrium}} - m_{0\,\mathrm{min}})$$

where N is the number of exchangeable amide sites, m is the experimentally determined centroid mass, $m_{0\,\mathrm{min}}$ is the averaged centroid mass of the 0-min exchange period control, and $m_{\mathrm{equilibrium}}$ is the averaged centroid mass of the equilibrium exchange control.

For 0-min exchange-period control experiments, a 4-µl aliquot of SDS-unfolded bacteriorhodopsin prepared as described above was added to a chilled mixture of 36 µl of DX buffer, 5 µl of 2.0 M potassium phosphate pH 2.4 and 5 µl of 5% (w/v) acid-labile surfactant at 0 °C, immediately vortex-mixed for 5 s and flash-frozen until proteolysis and LC–MS analysis were performed as described above.

For equilibrium exchange control experiments, peptic peptides of each protein were initially prepared, exchanged to the equilibrium in 90% D_2O DX exchange buffer, and mass-analysed after adjusting the pH. In brief, peptic peptides were prepared by first incubating a 4-µl aliquot of SDS-unfolded bacteriorhodopsin in 10 mM sodium phosphate in D_2O at an uncorrected pH of 6.8, then quenching by adding 5 µl of 2.0 M potassium phosphate pH 2.4 in 90% (v/v) D_2O and 5 µl of 5% (w/v) acid-labile surfactant in 90% (v/v) D_2O, followed by proteolysis using immobilized pepsin, prewashed in 200 mM potassium phosphate pH 2.4 in 90% D_2O at room temperature, in a spin-filter for 10 min. The peptic peptides in the filtrate were then collected by centrifugation and the pH was adjusted to 6.85 by the addition of 4 µl of 2.0 M NaOH in 90% D_2O to 50 µl of filtrate, to facilitate the equilibrium exchange reaction, and incubated at 42 °C overnight. The pH of the equilibrium reaction was then readjusted to 2.5 by the addition of 2.5 µl of 3.3 M HCl in 90% D_2O at 22 °C to reproduce the minimized back exchange. An aliquot of 40 µl of each equilibrium sample was then transferred to a vial containing 5 µl of water and 5 µl of 2 M potassium phosphate pH 2.4, mixed for 5 s, flash-frozen in liquid N_2 and kept at −80 °C until MS analysis. Although the equilibrium controls were subject to pre-proteolysis, the frozen equilibrium control samples were again incubated in immobilized pepsin before LC–MS analysis just as the deuterium exchange samples as described above, to reproduce the back-exchange process during proteolysis.

DXMS software[32] was used to identify the peptic peptides in the LC–MS–MS data collected before deuterium exchange experiments. MagTran software[33] was used to calculate the centroid mass of the isotopic envelope identified by DXMS software. Laplace software[34] was used to transform the observed exchange kinetics into rate constant distributions by the maximum-entropy method protocol. For visualization purposes a weighted average rate, k_{av}, was calculated for each fragment from the corresponding rate constant distribution:

$$k_{av} = \int kA(k)\mathrm{d}k$$

in which k is the rate constant and $A(k)$ is the normalized abundance of amide sites exchanging with rate k.

The exchange time courses were determined in triplicate, providing the mean and standard deviation for each time point. To estimate errors for k_{av} values we simulated ten time-course data sets in which the simulated data points were generated by using the mean and standard deviation for each time point as found in the experimental time-course data. The standard deviation of the k_{av} values for the simulated data were then used as error estimates.

Analysis of unsatisfied hydrogen bonds. All helical membrane proteins solved as of October 2007 at a resolution of 1.7 Å or better were identified. If two proteins had more than 40% sequence identity, the lower-resolution structure was rejected, leaving a set of six protein structures (PDB IDs 1C3W (ref. 35), 1KQF (ref. 36), 1U7G (ref. 37), 2A65 (ref. 38), 2B2H (ref. 39) and 2F2B (ref. 40)). The central hydrophobic region of each protein was identified as described[28] by finding the most hydrophobic 30-Å slice of the structure perpendicular to the membrane normal. The membrane normal was taken as the biological oligomeric symmetry axis in each structure. Hydrogen-bond satisfaction was calculated with HBPLUS, using the relaxed criteria described in ref. 29 to obtain most unlikely unsatisfied donors and acceptors[34]. Alternative Asn, Gln and His orientations were explored and explicit water was included in the analysis to identify all possible hydrogen bonds.

Distance analysis of hydrogen bonds in membrane and soluble proteins. The distribution of hydrogen-acceptor distances of α-helix backbone–backbone hydrogen bonds in the hydrocarbon cores of same unique high-resolution membrane protein structures described above was compared with that of the hydrogen bonds in soluble proteins. For soluble proteins, a 30% sequence

identity cutoff returned a total of 839 structures in the same resolution range as each of the membrane proteins used in analysis. A hydrogen-acceptor distance of the α-helix backbone–backbone hydrogen bond was calculated with HBPLUS by detecting the hydrogen bond mediated by carbonyl acceptor at position i and amide donor at $i + 4$ in the helical backbone identified in the header of the PDB file by using the same relaxed criteria described above. For membrane proteins, the solvent accessibility of residues was calculated in the context of the biologically relevant oligomeric form. For buried backbone hydrogen bonds, we selected hydrogen bonds occurring between donor and acceptor residues in which both of the side chains were at least 90% buried. For surface hydrogen bonds, both donor and acceptor side-chain solvent accessibilities were 20% or higher. Student's t-test performed between the distributions of hydrogen-acceptor distances identified in the lowest-resolution soluble-protein structures (1.68 Å) and in the highest-resolution soluble-protein structures (1.40 Å) indicated that the effect of the variation in resolution on the hydrogen-acceptor distance distribution was negligible at the resolution range used in analysis (results not shown).

30. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
31. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **54,** 905–921 (1998).
32. Hamuro, Y. *et al.* Dynamics of cAPK type IIβ activation revealed by enhanced amide H/2H exchange mass spectrometry (DXMS). *J. Mol. Biol.* **327,** 1065–1076 (2003).
33. Zhang, Z. & Marshall, A. G. A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *J. Am. Soc. Mass Spectrom.* **9,** 225–233 (1998).
34. Zhang, Z., Li, W., Logan, T. M., Li, M. & Marshall, A. G. Human recombinant [C22A] FK506-binding protein amide hydrogen exchange rates from mass spectrometry match and extend those from NMR. *Protein Sci.* **6,** 2203–2217 (1997).
35. Luecke, H., Schobert, B., Richter, H. T., Cartailler, J. P. & Lanyi, J. K. Structure of bacteriorhodopsin at 1.55 Å resolution. *J. Mol. Biol.* **291,** 899–911 (1999).
36. Jormakka, M., Tornroth, S., Byrne, B. & Iwata, S. Molecular basis of proton motive force generation: structure of formate dehydrogenase-N. *Science* **295,** 1863–1868 (2002).
37. Khademi, S. *et al.* Mechanism of ammonia transport by Amt/MEP/Rh: structure of AmtB at 1.35 Å. *Science* **305,** 1587–1594 (2004).
38. Yamashita, A., Singh, S. K., Kawate, T., Jin, Y. & Gouaux, E. Crystal structure of a bacterial homologue of Na^+/Cl^--dependent neurotransmitter transporters. *Nature* **437,** 215–223 (2005).
39. Andrade, S. L., Dickmanns, A., Ficner, R. & Einsle, O. Crystal structure of the archaeal ammonium transporter Amt-1 from *Archaeoglobus fulgidus*. *Proc. Natl Acad. Sci. USA* **102,** 14994–14999 (2005).
40. Lees, A. M., Deconinck, A. E., Campbell, B. D. & Lees, R. S. Atherin: a newly identified, lesion-specific, LDL-binding protein in human atherosclerosis. *Atherosclerosis* **182,** 219–230 (2005).

LETTERS

Structural basis for EGFR ligand sequestration by Argos

Daryl E. Klein[1], Steven E. Stayrook[1], Fumin Shi[1], Kartik Narayan[1] & Mark A. Lemmon[1]

Members of the epidermal growth factor receptor (EGFR) or ErbB/HER family and their activating ligands are essential regulators of diverse developmental processes[1,2]. Inappropriate activation of these receptors is a key feature of many human cancers[3], and its reversal is an important clinical goal. A natural secreted antagonist of EGFR signalling, called Argos, was identified in *Drosophila*[4]. We showed previously that Argos functions by directly binding (and sequestering) growth factor ligands that activate EGFR[5]. Here we describe the 1.6-Å resolution crystal structure of Argos bound to an EGFR ligand. Contrary to expectations[4,6], Argos contains no EGF-like domain. Instead, a trio of closely related domains (resembling a three-finger toxin fold[7]) form a clamp-like structure around the bound EGF ligand. Although structurally unrelated to the receptor, Argos mimics EGFR by using a bipartite binding surface to entrap EGF. The individual Argos domains share unexpected structural similarities with the extracellular ligand-binding regions of transforming growth factor-β family receptors[8]. The three-domain clamp of Argos also resembles the urokinase-type plasminogen activator (uPA) receptor, which uses a similar mechanism to engulf the EGF-like module of uPA[9]. Our results indicate that undiscovered mammalian counterparts of Argos may exist among other poorly characterized structural homologues. In addition, the structures presented here define requirements for the design of artificial EGF-sequestering proteins that would be valuable anti-cancer therapeutics.

The 419-residue mature *Drosophila melanogaster* Argos was modified to improve crystallization by removing a poorly conserved region with multiple O-linked glycosylation sites (residues 1–87). In addition, a non-conserved 120-residue insert (residues 140–259) exclusive to drosophilid Argos molecules was replaced with the corresponding five amino acids (PDGRT) from *Apis mellifera* Argos (Supplementary Fig. 1). The resulting 217-residue protein ($Argos_{217}$) has greatly improved stability and purification properties, and binds Spitz with an affinity (K_d = 7.7 nM; Supplementary Fig. 2) similar to that previously measured[5] for full-length $Argos_{419}$ (K_d = 20 nM). Triclinic crystals of $Argos_{217}$ bound to the Spitz EGF-domain ($Spitz_{EGF}$, residues 48–99) grew at neutral pH with two complexes per asymmetric unit, and diffracted to 1.6 Å resolution. The structure was solved by multiwavelength anomalous dispersion (MAD), using the halide soak method[10] (Supplementary Table 1). Representative regions of electron density are shown in Fig. 1a, b. Structures of uncomplexed $Argos_{217}$ and $Spitz_{EGF}$ were also determined (to 2.5 and 1.5 Å, respectively) by molecular replacement.

Argos consists of three separate disulphide-bonded β-sheet domains (domains 1–3) that bear no resemblance to EGF domains. This three-domain composition was not discerned in sequence analyses. The three domains of Argos engulf the bound $Spitz_{EGF}$ molecule with a structure that is reminiscent of a C-clamp (Fig. 1c, d). Domains 2 and 3 constitute the 'jaws' of this clamp and make an intimate set of direct contacts with bound $Spitz_{EGF}$. Domain 1 forms the backbone of the C-clamp and does not contribute directly to ligand binding. $Spitz_{EGF}$ itself is very similar in structure to other known EGFR family ligands (Supplementary Fig. 3). Its three disulphide bonds generate three loops in the typical EGF domain structure, which are termed the A-loop, the B-loop and the C-loop (Fig. 1e). The Spitz B-loop protrudes into the crevice between domains 2 and 3 of Argos (Fig. 1c, d). The conformation of $Spitz_{EGF}$ is largely unaltered on binding to Argos, apart from small changes in the backbone at the B-loop tip (Supplementary Fig. 3) and the reorientation of certain interfacial side chains.

Domains 1, 2 and 3 of Argos superimpose remarkably well (with C_α root mean squared deviations of 1.3–1.9 Å) despite sharing little sequence identity (less than 30%). The overall architecture found in all three domains is shown in Fig. 2a. The cysteine residues form a disulphide-bonded core, from which two β-hairpins project to form a four-stranded β-sheet with a relatively unusual antiparallel 2143 topology[11]. The result is a flat domain that resembles part of a left hand, with the β-hairpins as two fingers plus a thumb-like projection emerging from the disulphide-bonded core. In domains 1 and 3 (but not domain 2), a knuckle-like protrusion also projects below the plane of the page in Fig. 2a. The positions of the C1–C3 and C4–C6 disulphides are almost identical in all three Argos domains (Fig. 2b), but the third (C2–C5) disulphide is missing from domain 2. The absence of this disulphide correlates with the lack of a knuckle in domain 2. Domain 1 is distinguished by the presence of a unique additional N-terminal β-strand (β1′) that is parallel to strand β2. The drosophilid-specific insertion in Argos occurs at the top of domain 1 in the orientation shown in Fig. 2, immediately before the knuckle (between C4 and C5). This insertion would probably project out and away from domain 1, with its ends constrained by the C2–C5 and C4–C6 disulphides.

Protein Data Bank searches with the DALI server[12] and a secondary-structure-matching algorithm[13] revealed that the three domains of Argos are significantly related to the three-finger toxin fold found in snake neurotoxins and cardiotoxins[7]—although the disulphide-bonding pattern is altered, and Argos has just two (rather than three) fingers per domain. Interestingly, the three-finger toxin fold is also found in the extracellular ligand-binding domains of receptors for transforming growth factor-β (TGF-β) family ligands[8]. As shown in Fig. 2c, two fingers of the extracellular ligand-binding domain of the type II activin receptor (ActRII)[14] overlay very well with domain 3 of Argos. Members of the Ly-6 superfamily also share this fold[7], including the receptor for uPA (ref. 9).

Argos 'clamps' $Spitz_{EGF}$ between domains 2 and 3 (Fig. 1c, d) and buries 35% of the ligand's surface. Domains 2 and 3 are roughly parallel to one another (Fig. 1c) and are stacked so that they present opposite surfaces to the $Spitz_{EGF}$ molecule sandwiched between them. With the hand analogy introduced above, domain 3 presents its palm to the bound ligand, whereas domain 2 contacts $Spitz_{EGF}$ using the back of the hand. The Spitz-binding regions on domains 2 and 3 can each be divided into two sites (Fig. 3): an A site and a B site.

[1]Department of Biochemistry and Biophysics, University of Pennsylvania School of Medicine, 809C Stellar-Chance Laboratories, 422 Curie Boulevard, Philadelphia, Pennsylvania 19104-6059, USA.

The A sites (d2A, d3A) lie on the β-sheet surfaces, and the B sites (d2B, d3B) involve the thumb and disulphide-bonded core. Site A on domain 2 (d2A) consists of a patch of hydrophobic side chains on the back-of-hand surface of fingers 1 and 2 (L301, L326, Y341 and F343), which makes van der Waals contacts with three aliphatic side chains from the Spitz B-loop (V72, I74 and V79). Site B on domain 2 (d2B) comprises a flat surface (cyan in Fig. 3) formed by side chains from the base of finger 2 (V323, Y325, S346 and P347) and the domain 2 thumb (F294). This flat surface packs against the carboxy-terminal part of $Spitz_{EGF}$, contacting M89, Q91 and Y95 in the C-loop. We previously identified S346 and P347 from site d2B as significant residues in a genetic screen for modifiers of an Argos misexpression phenotype in *Drosophila* eye development[15]. S346 and P347 are both conserved in all known Argos orthologues (Supplementary Fig. 1) and make direct hydrogen bonds with Spitz (Fig. 3). In domain 3, part of site A (d3A) involves polar side chains on the palm side of finger 1 (T363, R365, E373 and N375) that interact with the $Spitz_{EGF}$ B-loop. In addition, a cluster of hydrophobic side chains around the tip of finger 2 (F403, L404 and I408) contact residues from the $Spitz_{EGF}$ N terminus, A-loop and B-loop. A key feature of this interaction is the projection of the F403 and L404 side chains into a hydrophobic pocket on the Spitz surface formed by Y52, P55, P78 and Y80 (Fig. 3). In the B-site on domain 3 (d3B), side chains close to the base of the two fingers (Q357, P358, L361, N377 and S412) form a binding site for the Spitz A-loop helix (contacting Spitz T57, F58, W61 and Y62). Site d3B also accommodates the side chains of Spitz R92 and L64 (Fig. 3).

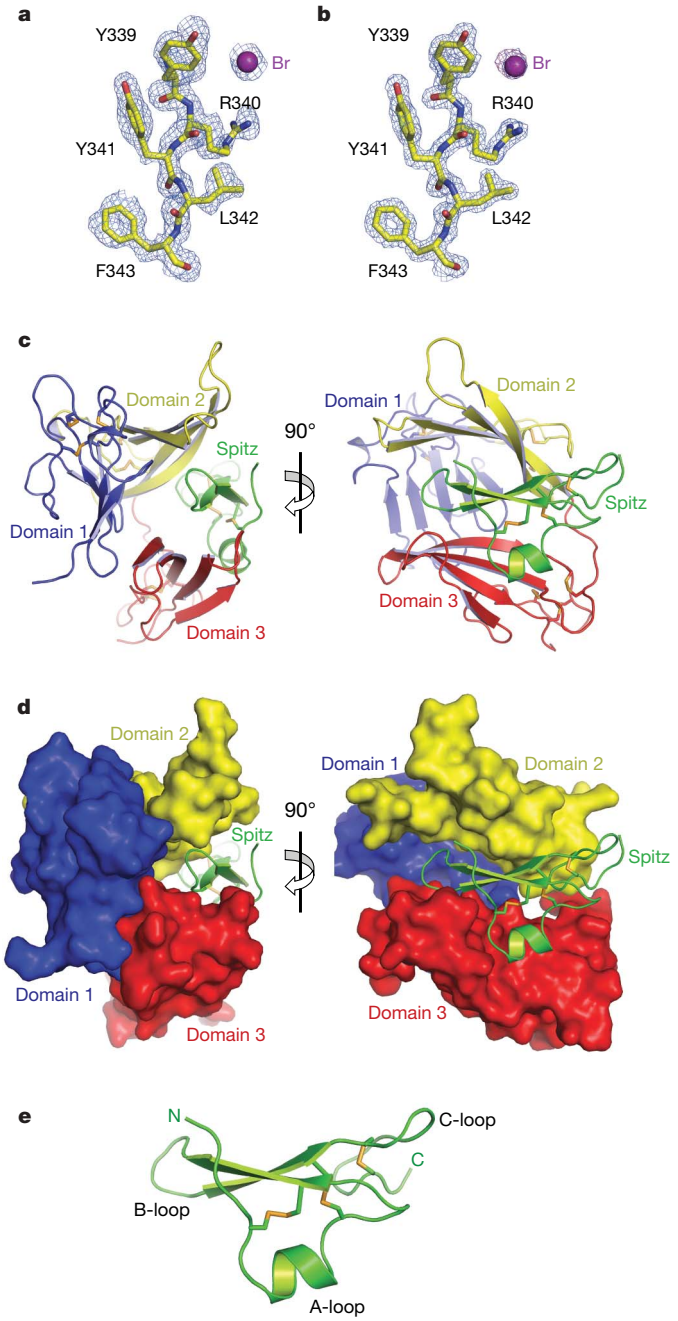


Figure 1 | Structure of the Argos–Spitz complex. **a**, Representative experimental electron density (contoured at 1σ) obtained after MAD phasing, showing a region of the Spitz-binding site on domain 2 (d2A; see Fig. 3). The initial model is shown placed in the density. **b**, The same region of an $2F_o - F_c$ map (contoured at 2σ) calculated using final phases. The final model is shown placed in the density. In purple, a peak corresponding to a bromide ion is seen in the anomalous difference Fourier map (contoured at 4σ) using Br peak data. **c**, Diagram of the $Argos_{217}$–$Spitz_{EGF}$ complex. Domains 1, 2 and 3 are coloured blue, yellow and red, respectively. Spitz is shown in green. Disulphide bridges are coloured orange. Two orthogonal views are shown. **d**, As in **c**, but with Argos in surface representation. **e**, Diagrammatic representation of the Spitz EGF domain structure, with A-loop, B-loop and C-loop marked.

Despite having a completely different structural scaffold from EGFR, Argos mimics the characteristic bipartite capture of growth factor seen in ligand-bound structures of the EGFR extracellular region (sEGFR)[16,17]. Argos presents two ligand-binding surfaces that closely resemble those in EGFR. Specifically, domain 2 of Argos mimics domain I of sEGFR in its ligand contacts, whereas domain 3 of Argos emulates sEGFR domain III (Fig. 4a). The primary ligand contacts made by domain 2 of Argos and domain I of EGFR are remarkably similar. Both use a central hydrophobic patch that interacts with a similar region of the B-loop of the bound EGF domain. As shown in Fig. 4a (upper panels), Argos site d2A and EGFR site 1 (from ref. 17) are very similar, with a comparable arrangement of hydrophobic side chains making ligand contacts in each case. The second binding site on Argos domain 2 (d2B) also recapitulates many other contacts between sEGFR domain I and human EGF (hEGF), but it is different in detail. Along the same lines, site B on Argos domain 3 (d3B) recapitulates site 2 in the sEGFR/hEGF interface[17] (Fig. 4a, lower panels), including interactions with an arginine residue that is critical for hEGF binding to its receptor (R41 in hEGF, R92 in Spitz). Both sites accommodate three key ligand side chains (Y13, L15 and R41 in hEGFR; Y62, L64 and R92 in Spitz) in analogous binding sites. Site 3 on domain III of EGFR is not mimicked by Argos; the C terminus of the bound EGF domain is much more exposed in the $Spitz_{EGF}$–Argos complex than in the hEGF–sEGFR complex (Fig. 4a). Argos compensates for the absence of these site 3 interactions with an extensive set of unique contacts mediated by site d3A (Fig. 4a).

Overall, Argos domain 2 buries slightly less surface on $Spitz_{EGF}$ (560 Å^2) than EGFR domain I buries on hEGF (745 Å^2)[17], but it includes more apolar surface (68%) than in the EGFR domain I/hEGF interface (56%). Argos domain 3 buries slightly more of its bound EGF domain (843 Å^2; 65% apolar) than does sEGFR domain III (819 Å^2; 62% apolar). Each interface has a shape complementarity parameter, S_c (ref. 18), of 0.70, which is typical for strong protein/protein interfaces and reflects a significantly greater shape complementarity than is seen in antibody/antigen interfaces. The fact that $Spitz_{EGF}$ binds about tenfold more strongly to Argos than to the *Drosophila* EGFR extracellular region may reflect different requirements for domain rearrangements in the two binding proteins. Clearly defined sets of (intramolecular) contacts between domains 1 and 2 and between domains 1 and 3 (Supplementary Fig. 4) may optimize the bipartite capture of $Spitz_{EGF}$ (or inhibit $Spitz_{EGF}$ dissociation). A crystal structure of unliganded $Argos_{217}$ (Supplementary Figs 5 and 6) suggests that domain 3 may be mobile, possibly 'collapsing' onto domain 2 in the absence of bound ligand (see Supplementary Information for discussion). Once $Spitz_{EGF}$ has

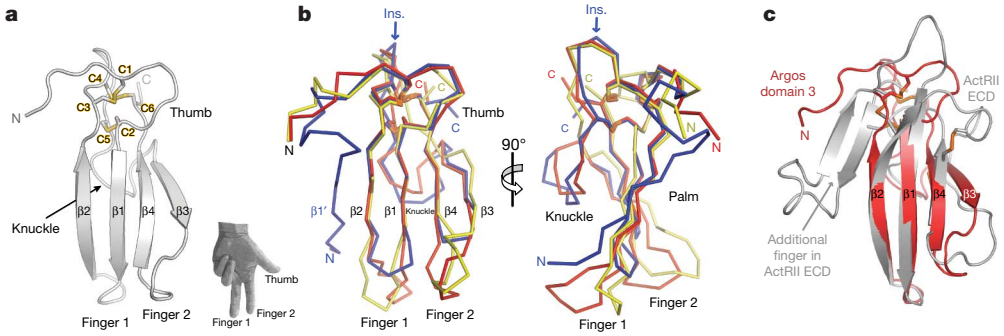


Figure 2 | Argos has three similar domains that resemble the three-finger toxin fold of TGF-β receptors. **a**, The overall fold of the three constituent domains in Argos is illustrated with domain 3. The four strands (β1–β4) form two fingers (fingers 1 and 2) that resemble those on a left hand (as shown). The 'palm' side of the domain faces out of the page. A knuckle-like protrusion projects below the page. At the top of the domain is a disulphide-bonded core from which emanate the two fingers plus the thumb (marked). Cysteine residues C1 to C6, which make C1–C3, C2–C5 and C4–C6 disulphides, are labelled, as are N and C termini. **b**, Domains 1, 2 and 3 are overlaid (as Cα ribbons) in the same orientation used in **a**. Colours are as in Fig. 1. Strand β1′, unique to domain 1, is labelled, as is the location of the 120 residue insert (Ins.) removed to generate $Argos_{217}$. Two orthogonal views are shown. **c**, Domain 3 of Argos (red) overlaid with the 100-residue extracellular ligand-binding domain (ECD) of the type II activin receptor (ActRII)[14] (coloured light grey; from Protein Data Bank code 2GOO).

bound, interactions between domains 1 and 3 in the complex may slow the dissociation of $Spitz_{EGF}$. About half of the residues involved in the binding of $Spitz_{EGF}$ to Argos are conserved in hEGF and/or human TGF-α, which may explain the ability of Argos to bind detectably (although weakly, with a K_d of about 5 μM) to hEGF (not shown).

Sequence analyses have failed to identify clear homologues of Argos in vertebrates, but this does not necessarily mean that functional mammalian analogues do not exist. The amino-acid sequence of Argos has been unusually cryptic, providing few (or misleading) clues about the structure of the protein. It only became apparent that Argos has three distinct domains (and no EGF-like domain) once the structure described here was determined. Moreover, the relationship of the constituent domains in Argos to the three-finger toxin fold can be seen only in structural (and not sequence) comparisons. As shown in Fig. 2c, the individual domains of Argos share unexpected and striking structural similarity with the extracellular ligand-binding regions of receptors for TGF-β/bone morphogenetic protein (BMP) family ligands, which consist of little more than a single three-finger toxin fold[8,14]. As shown in Supplementary Fig. 7, the positions of the ligand-binding sites in Argos domain 3 and the extracellular regions of TGF-β/BMP family receptors also correspond strikingly well. Both use the palm side of the domain as defined in the analogy drawn in Fig. 2a.

Among human proteins that contain three-finger toxin fold domains, one intriguing example uses three such modules to engulf

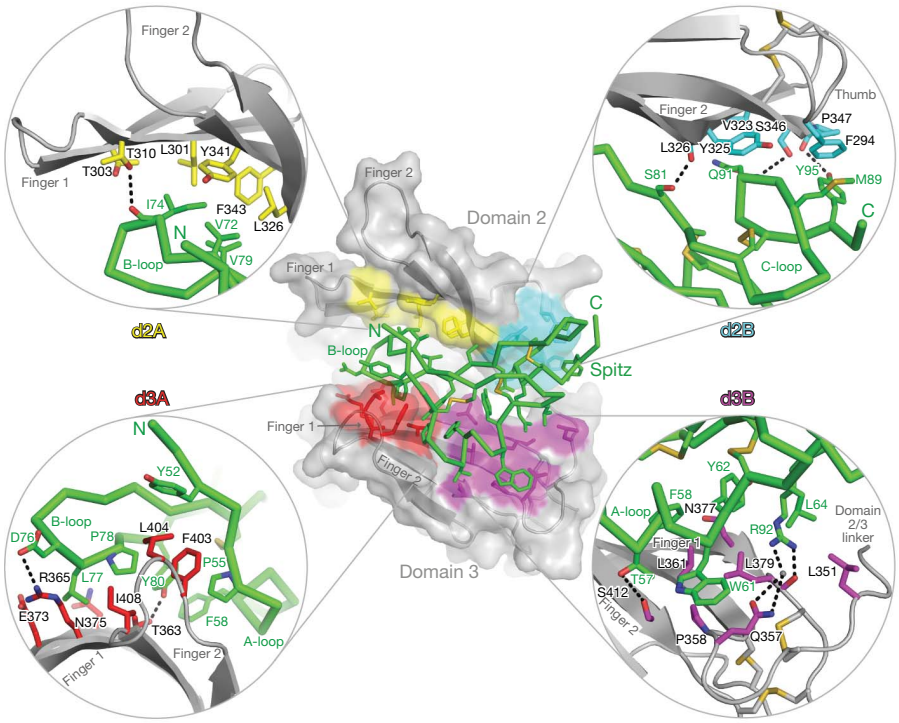


Figure 3 | Spitz-binding interactions. The centre panel shows the $Argos_{217}$–$Spitz_{EGF}$ complex in an orientation similar to that shown in Fig. 1d (right); Argos is coloured grey, Spitz green. Domains 2 and 3 are marked, as are their two fingers (which project to the left). Four individual Spitz-binding subsites are identified: d2A (yellow), d2B (cyan), d3A (red) and d3B (magenta). Surfaces of side chains involved in each subsite are coloured accordingly. In each of the four corners, details of an individual subsite are shown, with Argos side chains coloured for the site. Spitz is green in all panels, and the Argos backbone is grey.

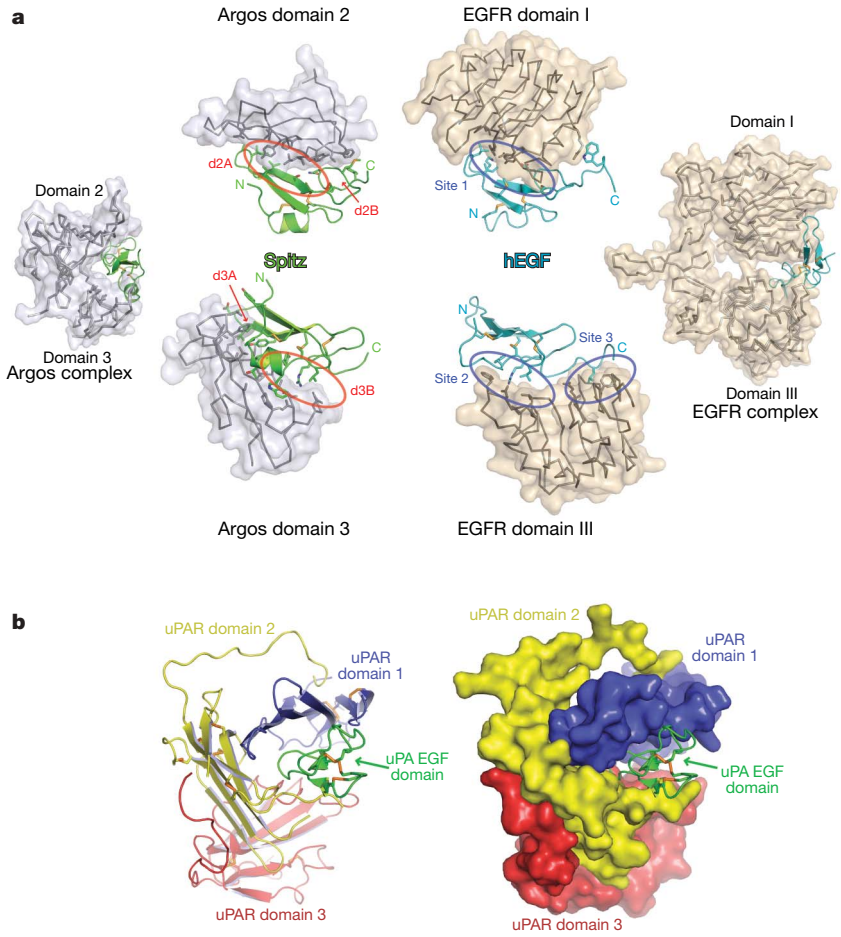


Figure 4 | Argos, EGFR and structural homologues entrap the EGF domain with two binding sites. **a**, The leftmost and rightmost panels show EGF domains bound to Argos and the human EGFR extracellular region[17] (sEGFR), respectively. Spitz is green and hEGF is cyan. In the central upper panels, $Spitz_{EGF}$ and hEGF are shown (in identical orientations) bound to Argos domain 2 (grey) and sEGFR domain I (beige). The side chains of EGF-domain-interacting residues are drawn. Site 1 on sEGFR domain I (defined in ref. 17) and its counterpart on Argos (which includes site d2A) are marked by blue and red ovals, respectively. In the lower central panels, $Spitz_{EGF}$ and hEGF (again in identical orientations) are shown bound to Argos domain 3 and sEGFR domain III. Sites 2 and 3 in the sEGFR/hEGF interface are marked with blue ovals. Argos site d3B mimics sEGFR site 2, but Argos does not mimic sEGFR site 3. Instead, Argos makes a unique set of interactions with $Spitz_{EGF}$ (site d3A). A key aliphatic side chain critical for the binding of hEGF to site 3 of EGFR (L47 in hEGF, I98 in Spitz) is disordered and exposed in the $Spitz_{EGF}$–Argos complex. **b**, Domain organization of uPAR[9,19]. The three domains in uPAR are coloured with the order used for Argos in Fig. 1. Like Argos, uPAR uses three copies of this domain type—although in a different arrangement—to form a C-clamp-like structure for enveloping an EGF domain[9,19].

an EGF-like domain. This is the cell-surface receptor for uPA[9,19], which is an unexpected structural homologue of Argos. As shown in Fig. 4b, the three domains from the uPA receptor (uPAR) form a clamp-like structure around the EGF domain found at the N terminus of uPA. Each domain presents its palm side to the bound ligand (like Argos domain 3). The uPAR structure resembles the clamp formed around the Spitz EGF domain by Argos (Fig. 1c, d). There are differences in the orientation of the bound EGF domain in the Argos–Spitz and uPAR–uPA complexes. Moreover, Argos has one of its three constituent domains (domain 2) 'inverted' so that it presents the back (rather than the palm) of the hand to the ligand. However, the correspondence in overall architecture and function (as proteins that entrap EGF domains) between uPAR and Argos suggests that other structural homologues of Argos should be sought in mammals. There are many human uPAR/Ly6 domain-containing proteins for which the function remains unclear. Several, such as CD177/PRV-1 and C4.4A, contain multiple three-finger domains[20,21] like uPAR. Moreover, C4.4A expression is known to be altered in several metastatic human cancers[22]. We suggest that one of these numerous structural homologues might represent a functional analogue of Argos. Even if such an analogue does not exist, the known human proteins from this class could clearly be used as structural scaffolds in the design of protein therapeutics that will sequester ErbB receptor-activating EGF domains.

It is increasingly clear that excessive or unregulated expression (or shedding) of ErbB-family ligands is important in numerous cancers, through autocrine and/or paracrine activation of cell growth[23–27]. The role of ErbB ligands may be particularly important in cancers for which available receptor-targeted approaches have failed or have met resistance[26]. In these (and other) cases, therapeutic agents that neutralize ErbB receptor ligands are likely to have great value. The understanding of Spitz neutralization by Argos that we present here provides avenues to explore in efforts to identify a human homologue of Argos. The structural lessons also provide clear suggestions for which human proteins might be used as scaffolds for generating protein therapeutics that sequester aberrantly produced EGF-like growth factors—exploiting a mechanism for inhibiting EGFR signalling that has evolved naturally.

METHODS SUMMARY

Protein purification and crystallization. $Argos_{217}$ was produced in baculovirus-infected *Spodoptera frugiperda* Sf9 cells, using the N-terminal BiP signal sequence to direct the secretion of the protein into the medium. The protein has a hexahistidine tag at its C terminus, which was used for purification

as described[5]. The EGF domain of Spitz (residues 48–99; $Spitz_{EGF}$) was generated by proteolytic cleavage of a modified form of secreted full-length Spitz produced in transfected *Drosophila* S2 cells. Crystals were grown using the hanging-drop method. Crystals grew from a 1:1 $Argos_{217}$:$Spitz_{EGF}$ mixture (250 μM complex) at pH 7, using PEG20000 as the precipitant, or from a 250 μM solution of $Argos_{217}$ alone using PEG3350 as precipitant (at pH 4.5). Crystals of $Spitz_{EGF}$ alone (500 μM) grew in ammonium sulphate, pH 6.5.

Structure determination. The $Argos_{217}$–$Spitz_{EGF}$ complex structure was determined by MAD with the halide soak method[10]. Crystallographic data were collected at the Advanced Photon Source and the Advanced Light Source, as summarized in Supplementary Table 1. Phasing from ten ordered bromide ions yielded a readily interpretable electron-density map allowing nearly the entire chain of each complex to be traced. Alternating cycles of model building with COOT[28] and refinement with REFMAC[29] led to a complete model of $Argos_{217}$ and $Spitz_{EGF}$ with R_{cryst} and R_{free} values of 0.20 and 0.24, respectively, to 1.6 Å resolution (Supplementary Table 1). The unliganded $Argos_{217}$ structure was solved by sequential molecular replacement using PHASER in the CCP4 suite of programs[29], and the $Spitz_{EGF}$ structure was solved by molecular replacement using a loop-truncated version of the hEGF domain structure (1JL9)[30].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 21 January; accepted 7 April 2008.
Published online 25 May 2008.

1. Holbro, T. & Hynes, N. E. ErbB receptors: directing key signaling networks throughout life. *Annu. Rev. Pharmacol. Toxicol.* **44,** 195–221 (2004).
2. Shilo, B. Z. Regulating the dynamics of EGF receptor signaling in space and time. *Development* **132,** 4017–4027 (2005).
3. Hynes, N. E. & Lane, H. A. ERBB receptors and cancer: the complexity of targeted inhibitors. *Nature Rev. Cancer* **5,** 341–354 (2005).
4. Freeman, M., Klambt, C., Goodman, C. S. & Rubin, G. M. The *argos* gene encodes a diffusible factor that regulates cell fate decisions in the *Drosophila* eye. *Cell* **69,** 963–975 (1992).
5. Klein, D. E., Nappi, V. M., Reeves, G. T., Shvartsman, S. Y. & Lemmon, M. A. Argos inhibits epidermal growth factor receptor signalling by ligand sequestration. *Nature* **430,** 1040–1044 (2004).
6. Kretzschmar, D. *et al. giant lens*, a gene involved in cell determination and axon guidance in the visual system of *Drosophila melanogaster*. *EMBO J.* **11,** 2531–2539 (1992).
7. Tsetlin, V. Snake venom α-neurotoxins and other 'three-finger' proteins. *Eur. J. Biochem.* **264,** 281–286 (1999).
8. Greenwald, J., Fischer, W. H., Vale, W. W. & Choe, S. Three-finger toxin fold for the extracellular ligand-binding domain of the type II activin receptor serine kinase. *Nature Struct. Biol.* **6,** 18–22 (1999).
9. Barinka, C. *et al.* Structural basis of interaction between urokinase-type plasminogen activator and its receptor. *J. Mol. Biol.* **363,** 482–495 (2006).
10. Dauter, Z., Dauter, M. & Rajashankar, K. R. Novel approach to phasing proteins: derivatization by short cryo-soaking with halides. *Acta Crystallogr. D Biol. Crystallogr.* **56,** 232–237 (2000).
11. Zhang, C. & Kim, S.-H. The anatomy of protein β-sheet topology. *J. Mol. Biol.* **299,** 1075–1089 (2000).
12. Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233,** 123–138 (1993).
13. Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D Biol. Crystallogr.* **60,** 2256–2268 (2004).
14. Allendorph, G. P., Vale, W. W. & Choe, S. Structure of the ternary signaling complex of a TGF-β superfamily member. *Proc. Natl Acad. Sci. USA* **103,** 7643–7648 (2006).
15. Alvarado, D., Evans, T. A., Sharma, R., Lemmon, M. A. & Duffy, J. B. Argos mutants define an affinity threshold for spitz inhibition *in vivo*. *J. Biol. Chem.* **281,** 28993–29001 (2006).
16. Garrett, T. P. J. *et al.* Crystal structure of a truncated epidermal growth factor receptor extracellular domain bound to transforming growth factor α. *Cell* **110,** 763–773 (2002).
17. Ogiso, H. *et al.* Crystal structure of the complex of human epidermal growth factor and receptor extracellular domains. *Cell* **110,** 775–787 (2002).
18. Lawrence, M. C. & Colman, P. M. Shape complementarity at protein/protein interfaces. *J. Mol. Biol.* **234,** 946–950 (1993).
19. Huai, Q. *et al.* Structure of human urokinase plasminogen activator in complex with its receptor. *Science* **311,** 656–659 (2006).
20. Rösel, M., Claas, C., Seiter, S., Herlevsen, M. & Zöller, M. Cloning and functional characterization of a new phosphatidyl-inositol anchored molecule of a metastasizing rat pancreatic tumor. *Oncogene* **17,** 1989–2002 (1998).
21. Temerinac, S. *et al.* Cloning of PRV-1, a novel member of the uPAR receptor superfamily, which is overexpressed in polycythemia rubra vera. *Blood* **95,** 2569–2576 (2000).
22. Hansen, L. V., Laerum, O. D., Illemann, M., Nielsen, B. S. & Ploug, M. Altered expression of the urokinase receptor homologue, C4.4A, in invasive areas of human esophageal squamous cell carcinoma. *Int. J. Cancer* **122,** 734–741 (2008).
23. Kenny, P. A. & Bissell, M. J. Targeting TACE-dependent EGFR ligand shedding in breast cancer. *J. Clin. Invest.* **117,** 337–345 (2007).
24. Zhou, B. B. *et al.* Targeting ADAM-mediated ligand cleavage to inhibit HER3 and EGFR pathways in non-small cell lung cancer. *Cancer Cell* **10,** 39–50 (2006).
25. Fujimoto, N. *et al.* High expression of ErbB family members and their ligands in lung adenocarcinomas that are sensitive to inhibition of epidermal growth factor receptor. *Cancer Res.* **65,** 11478–11485 (2005).
26. Hynes, N. E. & Schlange, T. Targeting ADAMS and ERBBs in lung cancer. *Cancer Cell* **10,** 7–11 (2006).
27. Borrell-Pagès, M., Rojo, F., Albanell, J., Baselga, J. & Arribas, J. TACE is required for the activation of the EGFR by TGF-α in tumors. *EMBO J.* **22,** 1114–1124 (2003).
28. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60,** 2126–2132 (2004).
29. CCP4 (Collaborative Computational Project Number 4). The CCP4 suite: Programs for protein crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **50,** 760–763 (1994).
30. Lu, H. S. *et al.* Crystal structure of human epidermal growth factor and its dimerization. *J. Biol. Chem.* **276,** 34913–34917 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank members of the Lemmon and Ferguson laboratories, G. Van Duyne and J. Shorter for advice and critical reading of the manuscript. This work was supported by grants from the National Institutes of Health (to M.A.L.) and the US Army Breast Cancer Research Program (to D.E.K. and M.A.L.).

Author Contributions D.E.K. and M.A.L. conceived and designed the project. D.E.K. was responsible for all construct design and execution of protein biochemistry, crystallization, and data collection. D.E.K. solved and refined the $Argos_{217}$–$Spitz_{EGF}$ complex structure. S.E.S. solved and refined the structures of uncomplexed $Argos_{217}$ and $Spitz_{EGF}$ by molecular replacement using datasets collected by D.E.K. K.N. helped with crystal manipulation and data collection. F.S. performed binding studies with Argos and Spitz variants, as well as analytical ultracentrifugation, directed by D.E.K. D.E.K. and M.A.L. interpreted data and wrote the manuscript.

Author Information Coordinates have been deposited in the Protein Data Bank under codes 3CA7 ($Spitz_{EGF}$), 3C9A ($Argos_{217}$–$Spitz_{EGF}$ complex), and 3CGU ($Argos_{217}$ alone). Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.A.L. (mlemmon@mail.med.upenn.edu).

METHODS

Argos constructs. To establish the normal signal sequence cleavage site of *Drosophila melanogaster* Argos, the N terminus of the mature recombinant protein was sequenced. The N-terminal sequence was TRLPLEVF, indicating that mature Argos is a 419-residue secreted protein. The non-conserved N terminus of Argos has little predicted secondary structure and contains multiple O-linked glycosylation sites. Fusing a BiP signal sequence to R88 of mature Argos produced a well-behaved protein that did not seem to be O-glycosylated.

D. melanogaster Argos also contains a proteolytically labile 120-residue insertion of low conservation (compared to those of other drosophilids), and little predicted secondary structure, between the fourth and fifth cysteine residues of the protein. Non-drosophilid Argos homologues contain only a five-residue linker in this region. We replaced the 120-residue insertion of *D. melanogaster* Argos with the corresponding five-residue linker (PDGRT) found in *Apis mellifera* Argos (Supplementary Fig. 1). The resulting protein ($Argos_{217}$) was well expressed and was resistant to proteolytic degradation. It contains 217 amino-acid residues, corresponding to residues 88–139 of mature *D. melanogaster* Argos linked (through the PDGRT sequence) to residues 260–419. A hexahistidine tag was appended to the C terminus to aid purification.

Argos production and purification. $Argos_{217}$ used for crystallization of the $Argos_{217}$–$Spitz_{EGF}$ complex was produced by secretion from Sf9 (*Spodoptera frugiperda*) cells using the Bac-to-Bac baculovirus expression system (Invitrogen Inc.) in accordance with the manufacturers' recommendations. About three days after infection of cells with recombinant virus, conditioned Sf900II medium (Invitrogen-Gibco) was harvested and separated from cellular material by brief centrifugation. The medium was then passed over TALON resin (ClonTech Inc.) for immobilized metal affinity chromatography (IMAC). The column was washed with 3–6 volumes of 10 mM MES pH 6.3, 150 mM NaCl containing 50 mM imidazole. $Argos_{217}$ was subsequently eluted with 300 mM imidazole in the same buffer. The eluted protein (more than 90% pure by Coomassie staining) was directly loaded onto a cation exchange column (S2; Bio-Rad Inc.) in the same buffer, and eluted with a gradient of NaCl concentration ($Argos_{217}$ elutes at about 1 M NaCl). Immediately before crystallization or binding studies, $Argos_{217}$ was gel-filtered into 10 mM MES pH 6.3, 150 mM NaCl on a Superose 12 column (GE Healthcare).

Crystals of unliganded $Argos_{217}$ were obtained with protein produced from *Drosophila* Schneider 2 (S2) cells as described previously[5], and purified exactly as described above. This protein behaves identically in all respects to $Argos_{217}$ produced by Sf9 cells. Biosensor studies (Supplementary Fig. 2) established that histidine-tagged $Argos_{217}$ binds to $Spitz_{EGF}$ with the same affinity as reported for wild-type $Argos_{419}$ in our previous studies[5].

Production of Spitz EGF domain. The coding region for the *D. melanogaster* Spitz extracellular region (ending at residue 99) was subcloned into the S2 cell expression vector pMT/BiP/V5-HisA (Invitrogen) so that the sequence RHHHHHHSMSGT immediately follows the BiP signal sequence cleavage site. The first serine in this sequence corresponds to S16 of mature secreted Spitz. A Factor Xa cleavage site was also engineered between residues 47 and 48 of secreted Spitz ($N_{46}I_{47}$TIEGR/$T_{48}F_{49}P_{50}$), where T_{48} represents the first residue of the EGF domain. Cleavage with Factor Xa allows removal of the highly glycosylated Spitz N terminus. In addition, deletion of the N-terminal 15 residues avoids lipid modification of the first cysteine[31] and substantially increases protein yield. S2 cells that stably express this modified form of secreted Spitz were selected by co-transfection with pCo-PURO and selection with puromycin[32], and the secreted protein was purified exactly as described[5]. After purification, the protein was cleaved with Factor Xa and the 52-residue EGF domain of Spitz (T48–D99; $Spitz_{EGF}$) was isolated by size-exclusion chromatography on a Superdex Peptide column (GE Healthcare) and the N-terminal fragment (plus uncleaved protein) was removed with IMAC. $Spitz_{EGF}$ binds $Argos_{419}$ and $Argos_{217}$ with the same affinity as the intact secreted form of Spitz[5].

Crystallization. $Argos_{217}$–$Spitz_{EGF}$ complex crystals grew from 0.1 M HEPES pH 7.0 and 24% ethylene glycol at 21 °C, with the addition of low concentrations (0.1–1%) of PEG20000 to slow crystal growth and thus improve crystal size and quality. Brief manipulation freed single crystal fragments that grew further over 7 days and were subsequently frozen directly in liquid nitrogen. Maximum single crystal dimensions reached 150 μm × 100 μm × 50 μm. Crystals were of space group $P1$, with unit cell dimensions $a = 50.0$ Å, $b = 51.3$ Å, $c = 70.0$ Å and $\alpha = 84.2°$, $\beta = 74.8°$, $\gamma = 75.7°$. There are two complexes per asymmetric unit, with a Matthews coefficient of 2.6 $Å^3\,Da^{-1}$, giving a solvent content of 53%.

Crystals of uncomplexed $Spitz_{EGF}$ grew from 15 mM ammonium sulphate in 0.1 M MES pH 6.5 containing 24% ethylene glycol. Crystals grew as single rods over two weeks, and were frozen directly from the drop in liquid nitrogen. Crystals were of space group $C2$, with unit cell dimensions $a = 58.3$ Å, $b = 36.2$ Å, $c = 25.4$ Å and $\alpha = 90°$, $\beta = 103.1°$, $\gamma = 90°$. There is one molecule per asymmetric unit, with a Matthews coefficient of 2 $Å^3\,Da^{-1}$ and a solvent content of 39%.

Crystals of unliganded $Argos_{217}$ grew at 18 °C from 10–20% PEG3350, 0.1 M sodium acetate pH 4.5, containing 0.2 M ammonium sulphate. Crystals were rapidly passed through paraffin oil for freezing. Crystals were of space group $C2$ with unit cell dimensions $a = 113.6$ Å, $b = 64.2$ Å, $c = 72.5$ Å and $\alpha = 90°$, $\beta = 101.6°$, $\gamma = 90°$. There are two molecules per asymmetric unit, with a Matthews coefficient of 2.5 $Å^3\,Da^{-1}$ and a solvent content of 52%.

Structure determination. For experimental phasing, efforts to introduce a variety of anomalous scatterers were made. Halide soaks[10] were a focus because they have been successful for several other disulphide-rich glycoproteins with few reactive side chains. Immediately before freezing, 1 M NaBr (in 5% PEG20000, 0.1 M HEPES pH 7.0, 24% ethylene glycol) was directly added (1:1) to the $Argos_{217}$–$Spitz_{EGF}$ complex crystal drops. A three-wavelength MAD data set was collected on a single NaBr-soaked crystal at Advanced Photon Source 23-IDD. Data on a second NaBr-soaked crystal were collected at a fourth wavelength with less attenuation for higher-resolution data. Data were processed with HKL2000 (ref. 33) and the phases were determined with SHELX C/D/E[34,35], using all four data sets and anomalous signal from ten bromide ions. The resulting electron density map was readily interpretable, allowing almost the entire chain of each complex in the asymmetric unit to be traced straightforwardly. Alternating cycles of model building with COOT[28] and refinement with REFMAC[29] led to a complete model of $Argos_{217}$ and $Spitz_{EGF}$. The first ten residues in both $Argos_{217}$ molecules are not seen in the crystal structure; nor are the C-terminal hexahistidine tags. In addition, the first and last two residues in $Spitz_{EGF}$ could not be located in the complex. NCS averaging was used for initial rounds of refinement but released in the final stages of refinement.

The structure of unliganded $Argos_{217}$ was solved by sequential molecular replacement using PHASER[36] in the CCP4 program suite[29]. Domains 1 and 2 of $Argos_{217}$ from the complex were used to find a molecular replacement solution, and a solution was then identified for domain 3. The structure of $Spitz_{EGF}$ was solved by molecular replacement by using a loop-truncated version of the hEGF domain structure (1JL9)[30].

Calculations and figure preparation. Calculations of buried surface were performed with AREAIMOL in the CCP4 suite of programs[29]. Calculations of surface complementarity, S_c (ref. 18), used the program SC in CCP4 (ref. 29). Quantitative descriptions of protein domain movement were calculated with the DynDom server[37]. Structure validation was performed with SFCHECK and PROCHECK in CCP4 (ref. 29). Figures were prepared with PyMOL[38].

31. Miura, G. I. *et al.* Palmitoylation of the EGFR ligand Spitz by Rasp increases Spitz activity by restricting its diffusion. *Dev. Cell* **10,** 167–176 (2006).
32. Iwaki, T., Figuera, M., Ploplis, V. A. & Castellino, F. J. Rapid selection of *Drosophila* S2 cells with the puromycin resistance gene. *Biotechniques* **35,** 482–486 (2003).
33. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
34. Pape, T. & Schneider, T. R. HKL2MAP: a graphical user interface for phasing with SHELX programs. *J. Appl. Cryst.* **37,** 843–844 (2004).
35. Schneider, T. R. & Sheldrick, G. M. Substructure solution with SHELXD. *Acta Crystallogr. D Biol. Crystallogr.* **58,** 1772–1779 (2002).
36. McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C. & Read, R. J. Likelihood-enhanced fast translation functions. *Acta Crystallogr. D Biol. Crystallogr.* **61,** 458–464 (2005).
37. Hayward, S. & Lee, R. A. Improvements in the analysis of domain motions in proteins from conformational change: DynDom version 1.50. *J. Mol. Graph. Model.* **21,** 181–183 (2002).
38. DeLano, W. L. *The PyMOL Molecular Graphics System* (DeLano Scientific, Palo Alto, CA, 2002).

LETTERS

Translation factors promote the formation of two states of the closed-loop mRNP

Nadia Amrani[1], Shubhendu Ghosh[1], David A. Mangus[1] & Allan Jacobson[1]

Efficient translation initiation and optimal stability of most eukaryotic messenger RNAs depends on the formation of a closed-loop structure and the resulting synergistic interplay between the 5′ m^7G cap and the 3′ poly(A) tail[1,2]. Evidence of eIF4G and Pab1 interaction supports the notion of a closed-loop mRNP[3], but the mechanistic events that lead to its formation and maintenance are still unknown. Here we use toeprinting and polysome profiling assays to delineate ribosome positioning at initiator AUG codons and ribosome–mRNA association, respectively, and find that two distinct stable (resistant to cap analogue) closed-loop structures are formed during initiation in yeast cell-free extracts. The integrity of both forms requires the mRNA cap and poly(A) tail, as well as eIF4E, eIF4G, Pab1 and eIF3, and is dependent on the length of both the mRNA and the poly(A) tail. Formation of the first structure requires the 48S ribosomal complex, whereas the second requires an 80S ribosome and the termination factors eRF3/Sup35 and eRF1/Sup45. The involvement of the termination factors is independent of a termination event.

In vitro translation reactions used synthetic mRNAs derived from yeast transcripts, extracts that recapitulate synergy between the cap and the poly(A) tail (Supplementary Fig. 1a), competitive inhibition of translation initiation by m^7GpppG (cap analogue), and analyses of ribosome positioning or mRNA association by toeprinting and sucrose-gradient sedimentation. Addition of the elongation inhibitor cycloheximide to translation reactions programmed by the 2,135-nucleotide (nt) AAA and UAA mRNAs containing long or short open reading frames, respectively (Fig. 1a), allowed the detection of cycloheximide-dependent initiator AUG toeprints that reflect 80S ribosomes protecting 16–18 nt 3′ of the AUG codon[4–6] (Fig. 1b, left and top right panels, lanes 1 and 3). These toeprints were dependent on initiation codon recognition, the presence of yeast extract, and concurrent mRNA translation (Supplementary Fig. 1b, c). Supporting the latter conclusion, toeprints were almost completely eliminated by 2.7 mM cap analogue, a concentration that distinguished genuine toeprints from background bands (Fig. 1b, left and top right panels, lanes 2 and 4). Lower concentrations of cap analogue also inhibited AUG toeprint accumulation, with 70% and 96% sensitivity obtained at 0.05 mM and 0.5 mM, respectively (Fig. 1b, bottom right panel). A shorter mRNA (miniUAA1, 488 nt; Fig. 1a) also yielded the AUG toeprint (Fig. 1c, left panel, lane 1), but this band was resistant to 2.7 mM cap analogue (lane 2) and manifested sensitivity only at higher concentrations (Fig. 1c, right panel). Thus, in wild-type extracts, the short capped (see Supplementary Fig. 1d) and polyadenylated miniUAA1 mRNA is about 160-fold more resistant to cap analogue than the longer AAA mRNA. The miniADE2 (485 nt) and ADE2 (2,070 nt) mRNAs, whose respective sizes (but not sequences) are comparable to those of the miniUAA1 and AAA transcripts, also show the same results (Supplementary Fig. 2a). In agreement with the apparent mRNA size dependence of cap analogue resistance *in vitro*, mRNAs with intermediate sizes (from 488 to 2,135 nt), but with the same stability (Supplementary Fig. 2b), showed intermediate phenotypes, with about 40% sensitivity to 2.7 mM cap analogue obtained with an 882-nt mRNA, 60% sensitivity with a 1,105-nt mRNA, and near-maximal sensitivity with a 1,336-nt mRNA (Fig. 1d).

To rule out cap-independent translation as the basis for cap analogue resistance, we translated and toeprinted a polyadenylated miniUAA1 mRNA with no 5′ cap. This mRNA, which is stable during translation (Supplementary Fig. 2c), initiates inefficiently in wild-type extracts (Fig. 2a, lane 1), strongly suggesting that translation of our reporter mRNAs is cap-dependent. As reported previously[7,8], the addition of cap analogue stimulates the translation of uncapped mRNA (Fig. 2a, lane 2), possibly because it titrates other inhibitors. Taken together, these data suggest that interactions between the eIF4F cap-binding complex and the 5′ m^7G cap of the mini-mRNAs are much stronger than those occurring with the long mRNAs.

Short mRNAs may be preferentially resistant to cap analogue because they form a more stable closed-loop mRNP than long mRNAs, possibly by promoting increased affinity of mRNP factors for each other or for mRNA structures such as the 5′ cap[3,9–11]. To test the relationship of the closed-loop state to m^7GpppG resistance, we analysed the toeprinting of capped, poly(A)$^-$ miniUAA1 mRNA. This transcript showed strong sensitivity to cap analogue (Fig. 2b, lane 2). Further analyses of miniUAA1 mRNAs having different poly(A) tail lengths (but identical stabilities; Supplementary Fig. 2d) showed that miniUAA1 mRNA with a poly(A) tail of 14 or 18 adenosines (Fig. 2c, lanes 1–4) displayed a phenotype similar to poly(A)-deficient mRNA (more than 95% sensitivity to 2.7 mM cap analogue). Increasing the poly(A) tail length to 22, 25 or 33 residues yielded an intermediate phenotype (about 80% sensitivity to cap analogue; Fig. 2c, lanes 5–10), whereas a miniUAA1 mRNA with 57 adenosines showed no sensitivity to 2.7 mM cap analogue (Fig. 2c, lanes 11 and 12). These results correlate cap analogue resistance with poly(A) length, a result consistent with the formation of a closed loop. Because Pab1–poly(A) association requires a minimum of 12 adenosines, and multiple Pab1 molecules can bind the same poly(A) tract in a 27-nt repeating unit[12–14], it seems that at least two Pab1 molecules are required for a stable closed-loop structure. In concord with this conclusion, previous studies showed that an A_{15} tail did not suffice to stimulate translation in *Drosophila*[15] and mammalian[16] extracts, but that longer poly(A) tails promoted strong translational enhancement.

Toeprinting of miniUAA1 mRNA after translation in Pab1-deficient (*pab1Δpbp1Δ*) extracts reinforced the notion of a critical role for Pab1 in cap analogue resistance. Translation initiation in these extracts was highly sensitive to cap analogue (Fig. 3a, lanes 3 and 4; compare with extracts from wild-type (lanes 1 and 2) or *pbp1Δ* (a *pab1Δ* suppressor) cells (lanes 5 and 6)), and the apparent 80-fold

[1]Department of Molecular Genetics and Microbiology, University of Massachusetts Medical School, Worcester, Massachusetts 01655-0122, USA.

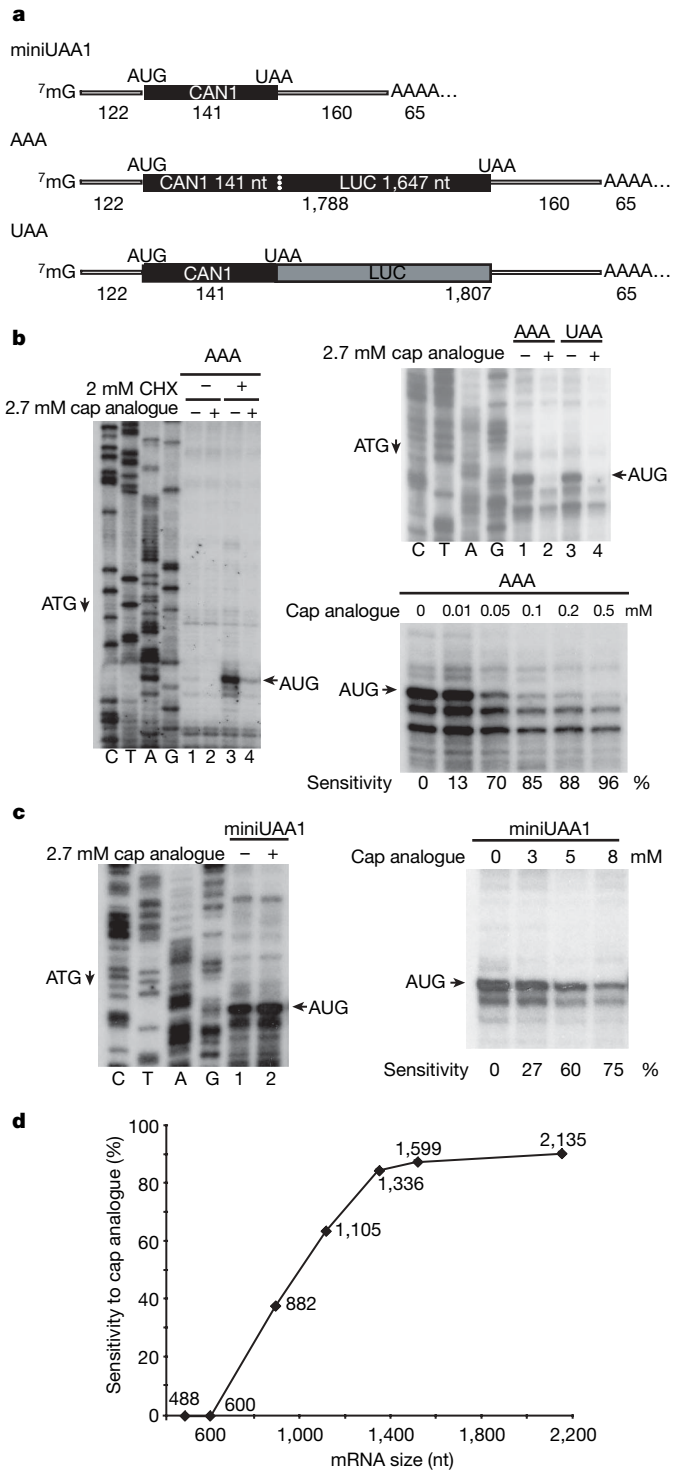


Figure 1 | Toeprint analyses of initiation on long and short mRNAs in the presence of cycloheximide in wild-type extracts. a, General diagram of the miniUAA1, AAA and UAA mRNAs. Sizes (in nucleotides) listed under each construct refer to the length of the 5′ UTR, the coding region, the 3′ UTR, and the poly(A) tail, respectively. **b**, Addition of cap analogue to AAA and UAA mRNAs inhibits the accumulation of the AUG toeprint bands. CHX, cycloheximide. **c**, miniUAA1 mRNA AUG toeprints are resistant to 2.7 mM cap analogue and become sensitive only at higher concentrations. **d**, Sensitivity to 2.7 mM cap analogue is dependent on mRNA size. The values on the graph are lengths in nucleotides and are averages of two independent experiments. Positions of the toeprints are indicated with arrows. The left portions of panels **b** and **c** show dideoxynucleotide sequencing reactions for the AAA or miniUAA1 templates (with 5′ to 3′ sequence reading from top to bottom).

increase in sensitivity to cap analogue (Fig. 1c, right panel, and Supplementary Fig. 5b, top panel) was directly attributable to the absence of Pab1 because supplementation with 15 pmol of recombinant Pab1, but not the same amount of BSA, restored cap analogue resistance to wild-type levels (Supplementary Fig. 5a). Although Pab1 is essential for cap analogue resistance, its level must be well balanced because excess Pab1—that is, the addition of 15 pmol of Pab1 to wild-type extracts (lanes 3 and 4) or 38 pmol of Pab1 to *pab1Δpbp1Δ* extracts (data not shown)—reduces the extent of cap analogue resistance.

To further evaluate the relationship between cap analogue resistance and the formation of a closed loop, we determined whether the same regions of Pab1 were required for both phenomena. Extracts derived from *pab1-134* cells, in which Pab1 has a wild-type affinity for eIF4G1 but a decreased affinity for eIF4G2 (ref. 17) (yeast has two eIF4G isoforms encoded by the functionally redundant *TIF4631* and *TIF4632* genes, respectively[18]), had a wild-type phenotype (Fig. 3a, lanes 7 and 8). However, toeprinting analyses in *pab1-184* extracts, in which Pab1 does not bind to eIF4G1 or eIF4G2 (ref. 17), revealed a strong sensitivity to cap analogue (Fig. 3a, lanes 9 and 10). Similarly, extracts of *pab1-ΔRRM1* cells, in which Pab1 has lost the ability to interact with the poly(A) tail, and probably with either isoform of eIF4G as well[19], also show strong sensitivity to cap analogue (Fig. 3a, lanes 11 and 12) but not to the same concentration of GTP (Supplementary Fig. 2e). Both *pab1-184* and *pab1-ΔRRM1* extracts show about tenfold more sensitivity to cap analogue than those obtained from cells lacking the *PAB1* gene (Supplementary Fig. 5b), possibly indicating dominant-negative effects of the mutant proteins. To determine whether Pab1 interaction with the termination factor eRF3/Sup35 (ref. 20) is also required for cap analogue resistance, we analysed extracts of *pab1-ΔC-term* cells. Figure 3a (lanes 13 and 14) shows that this mutation also confers sensitivity to cap analogue, although 2.5-fold less sensitivity than that observed in *pab1Δpbp1Δ* extracts (Supplementary Fig. 5b, d). In addition, the pattern of toeprint inhibition with the *pab1-ΔC-term* extract has a different appearance from that seen with other *pab1* mutants (Supplementary Fig. 5b, d), suggesting that the Pab1 carboxy-terminal domain may be involved in a step distinct from that involving the other domains. All of the *pab1* mutant extracts show full sensitivity to cap analogue when programmed with poly(A)$^-$ miniUAA1 mRNA (Supplementary Fig. 2f). Recombinant Pab1 could not complement the phenotypes of these extracts when supplemented with polyadenylated miniUAA1 mRNA, further suggesting dominant-negative effects of the mutated proteins (data not shown).

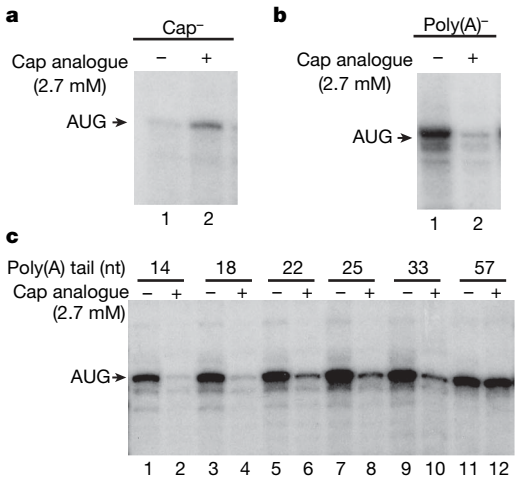


Figure 2 | Cap analogue resistance of the miniUAA1 mRNA is dependent on cap and poly(A) in wild-type extracts and suggests formation of a stable closed-loop structure. a, AUG toeprint analyses of uncapped mRNA. **b**, Poly(A)-deficient mRNAs are highly sensitive to cap analogue. **c**, Resistance to cap analogue is dependent on poly(A) size.

The differences in sensitivity of the different *pab1* extracts to cap analogue are consistent with the loss of interactions characteristic of the involvement of specific Pab1 domains[21] in closed-loop mRNP formation and stabilization. Accordingly, we analysed miniUAA1 mRNA toeprints in extracts from strains harbouring mutations in genes encoding different Pab1-interacting proteins. Extracts of *eIF4G1-ΔN300* cells[17], in which there is no eIF4G2 and Pab1–eIF4G1 interaction is disrupted, show strong sensitivity to cap analogue and correlate well with the cap analogue-sensitivity phenotype of *pab1-184* extracts (Fig. 3b and Supplementary Fig. 5c). Similarly, extracts harbouring mutated eRF3 (*sup35-R419G*)[22] or eRF1 (*sup45-2*)[23] termination factors show both sensitivity to cap analogue (but not to GTP; Supplementary Fig. 2e) and toeprint patterns that are strikingly similar to those obtained with the *pab1-ΔC-term* extract (Fig. 3c and Supplementary Fig. 5d), with the exception that they also yield two extra bands 5′ of the AUG toeprint that are suggestive of initiation anomalies (see Supplementary Data and Supplementary Fig. 3a). Control reactions demonstrated that the termination mutant extracts were fully sensitive to cap analogue when programmed with poly(A)$^-$ miniUAA1 mRNA (Supplementary Fig. 2g) and that the addition of extra Mg^{2+} to the different translation reactions (to compensate for a possible titration by cap analogue) had no effect on their toeprint phenotypes (Supplementary Fig. 2h). These results, and the observation that toeprinting of miniUAA1 mRNA after translation in extracts defective in eIF3 and eIF4E also exhibited more sensitivity to cap analogue than did wild-type extracts (Supplementary Fig. 4), imply that cap analogue resistance results from formation of a stable closed-loop structure.

Further insights into the formation of a closed loop followed from experiments using extracts supplemented with mutant transcripts or different competitive inhibitors, or independent analytical methods. First, we evaluated whether the role of eIF4G–Pab1–eRF3 interactions in the formation of a closed loop (and, possibly, ribosome recycling[24]) was also dependent on translation termination[24]. We prepared a mini-mRNA with no stop codon (nonstopminiUAA1) and analysed its response to cap analogue in extracts derived from wild-type or termination-defective cells. Translation of nonstopminiUAA1 in wild-type extracts showed strong cap analogue resistance (Fig. 3d, lanes 1 and 2), suggesting that conventional termination steps are not required for a stable closed-loop mRNP. In contrast, *sup35-R419G* and *sup45-2* extracts programmed with nonstopminiUAA1 RNA showed strong sensitivity to cap analogue (Fig. 3d, lanes 3–6), indicating that, although termination itself is not required, *de novo* formation of a stable closed-loop structure in yeast is dependent on the principal termination factors. Second, because there was no requirement for termination, we considered the possibility that the formation of a closed loop precedes the first round of mRNA translation; we therefore analysed the toeprinting of miniUAA1 mRNA after 2-min time courses of translation in wild-type or Pab1-mutant extracts. In wild-type extracts, cap analogue-resistant AUG toeprints were obtained within 30 s of incubation with miniUAA1 mRNA. Similarly, in *pab1-184* and *pab1-ΔRRM1* extracts, sensitivity to cap analogue was observed at the same time point (Fig. 3e, f). Third, to monitor mRNA–ribosome association by an independent method, we used sucrose-gradient fractionation to analyse the translation of miniUAA1 mRNA. Figure 3g shows that, in the absence of cap

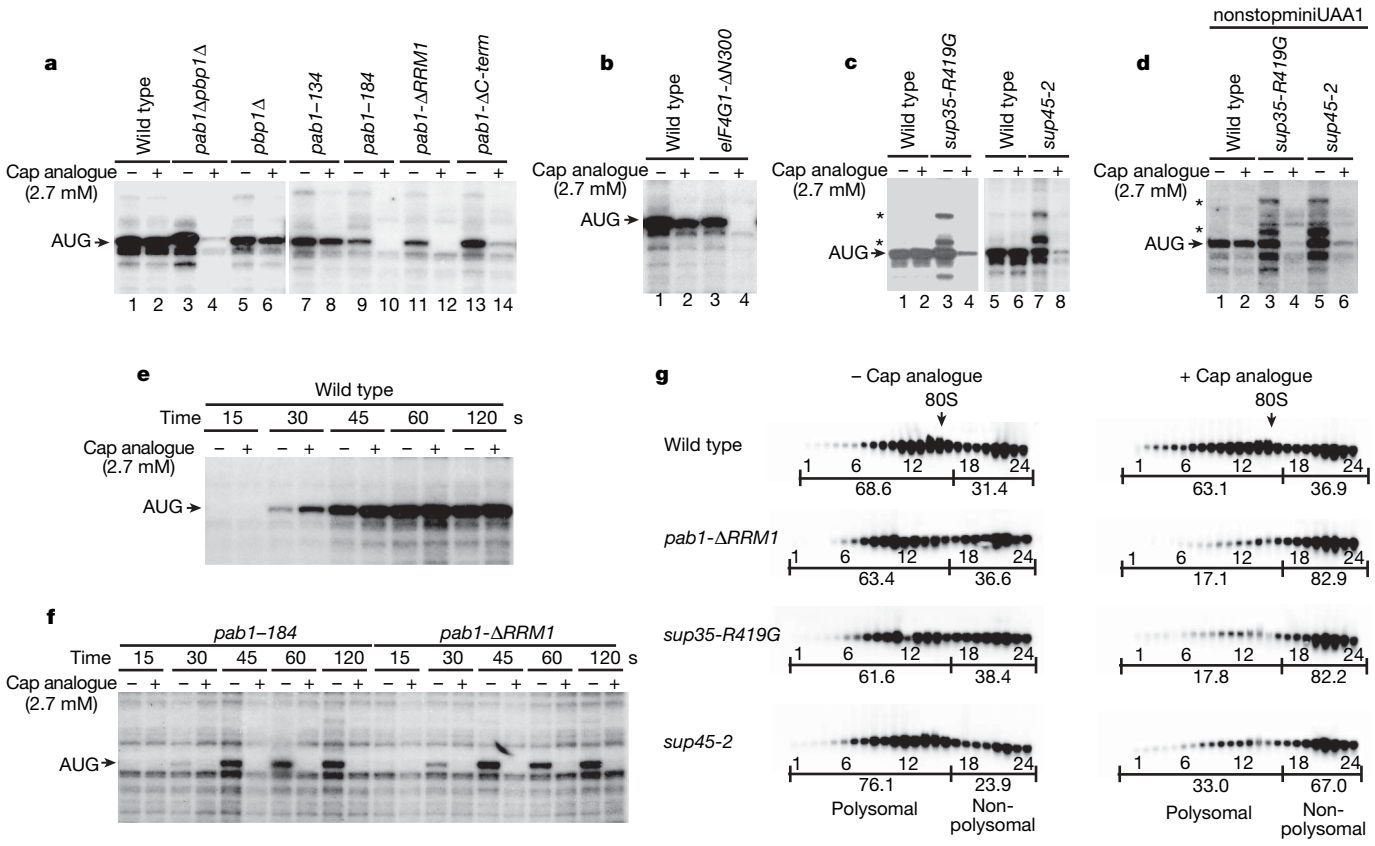


Figure 3 | Formation of a stable closed-loop structure on a capped and polyadenylated mRNA in the presence of an 80S complex requires Pab1 interactions with eIF4G, mRNA, and Sup35. a, Toeprinting analyses of miniUAA1 mRNA in wild-type, Pab1-defective or *pbp1Δ* extracts. **b**, Sensitivity to cap analogue in an eIF4G1 mutant incapable of Pab1 interaction. **c**, *sup35-R419G* and *sup45-2* mutants show sensitivity to cap analogue and additional toeprint bands upstream of the initiator AUG. **d**, Sensitivity to cap analogue is independent of the termination event. **e**, **f**, Cap analogue resistance or sensitivity of miniUAA1 mRNA appears at the onset of translation in wild-type (**e**) or Pab1-defective (**f**) extracts. **g**, Sucrose-gradient fractionation of miniUAA1 mRNA translated in wild-type and mutant extracts in the absence (left) and presence (right) of cap analogue, and in the presence of cycloheximide. In **g** the values above the horizontal lines depict fraction numbers and those below the lines denote the respective percentages of mRNA associated with polysomal and non-polysomal fractions.

analogue, 60–70% of miniUAA1 mRNA is associated with the polyribosomal fractions, whereas, in the presence of cap analogue, most of the miniUAA1 mRNA in the *pab1-ΔRRM1*, *sup35-R419G* and *sup45-2* extracts associated with the non-polysomal fractions. In contrast, the mRNA translated in the wild-type extract showed the same distribution as in the absence of the drug. Taken together, these data provide additional evidence that translation of capped and polyadenylated miniUAA1 mRNA is resistant to cap analogue in wild-type extracts.

The foregoing observations are consistent with the notion that interactions between the mRNP 5′ and 3′ ends are established early in the translation process (Fig. 3d–f). We therefore sought to identify the step of translation initiation associated with the formation of a closed loop. The toeprinting assays described above used cycloheximide to stabilize the 80S ribosome on the initiator AUG[4] and thus monitored a late step in initiation. To determine whether closed-loop mRNP formation occurred earlier, we analysed mRNA toeprints in extracts supplemented with the non-hydrolysable GTP analogue GMP-PNP[4], a drug that blocks the conversion of the 48S complex to an 80S ribosome *in vitro*. In both wild-type and *pbp1Δ* extracts, translation of miniUAA1 mRNA in the presence of GMP-PNP gave the same results as those observed with cycloheximide, namely full resistance to 2.7 mM cap analogue (Fig. 4a, lanes 1, 2 and 5, 6). This result implies that steps essential to establishing the closed-loop phenotype occur before or during formation of the 48S complex.

As observed with cycloheximide toeprinting, sensitivity of miniUAA1 mRNA in the presence of GMP-PNP was dependent on a poly(A) tail (data not shown) and the availability of functional Pab1 and eIF4G (Fig. 4a, lanes 3, 4, 7–16, and Fig. 4b, lanes 5–8). However, alterations in termination factor activity yielded different toeprint phenotypes. Extracts of termination factor mutants supplemented with GMP-PNP either at the beginning of the reaction (data not shown) or after 4 min of translation were fully resistant to cap analogue and lacked the upstream toeprints detected with cycloheximide (Fig. 4c). In agreement with these results, sucrose-gradient analyses of mRNA–ribosome association in extracts supplemented with GMP-PNP at the onset of the reaction showed no differences between the wild type and the termination mutant extracts in the presence or absence of cap analogue, whereas the *pab1-ΔRRM1* extract had a phenotype characteristic of inhibition of translation initiation by cap analogue (Supplementary Fig. 5e). Thus, although eRF1 and eRF3 do not seem to affect the rate of translation initiation (Supplementary Fig. 3d) or the formation of the closed-loop mRNP that includes only the 48S complex, they are required on formation of the 80S ribosome to generate a second state of the closed-loop structure.

The 5′ cap and 3′ poly(A) tail act synergistically to promote the stability and translatability of an mRNP[21]. Here we show that these mRNA appendages communicate as the first round of translation begins, establishing interactions involving at least two molecules of poly(A)-associated Pab1, the initiation factors eIF4G, eIF4E and eIF3, and the termination factors eRF1 and eRF3. Two forms of the stable closed-loop structure can be distinguished during initiation on short, but not long, mRNAs *in vitro*. The first requires the preinitiation 48S complex but not eRF1 and eRF3, whereas the second is formed after 60S joining and requires the two eRFs as well as all the other components of the first structure. The restriction of the efficient formation of a closed loop *in vitro* to short mRNAs suggests that other factors facilitating this process may be absent or inactive in our cell-free system.

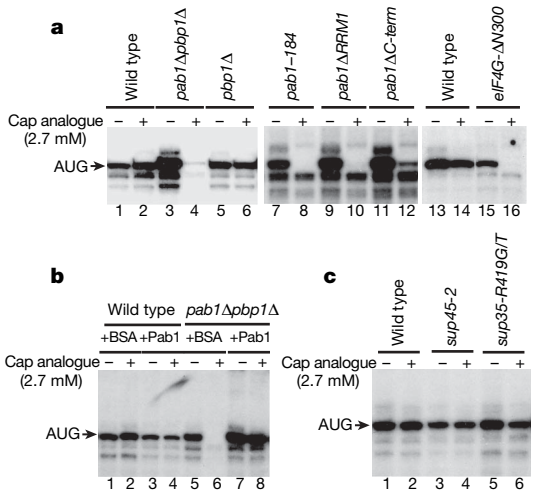


Figure 4 | Stabilization of the closed-loop structure on a capped and polyadenylated mRNA in the presence of a 48S complex requires interactions of Pab1 with eIF4G and mRNA. a, Extracts from Pab1-deficient and eIF4G1-deficient strains are sensitive to cap analogue in the presence of GMP-PNP. **b**, Addition of recombinant Pab1 restores resistance to cap analogue in *pab1Δpbp1Δ* extracts in the presence of GMP-PNP. **c**, Extracts from the *sup35-R419G* and *sup45-2* termination mutants are resistant to cap analogue in the presence of GTP analogue.

METHODS SUMMARY

Synthetic mRNAs of different sizes or sequences were translated in cell-free extracts derived from wild-type yeast cells or from mutant strains lacking the activity of specific translation factors. The elongation inhibitor cycloheximide or the non-hydrolysable GTP analogue GMP-PNP was used to trap 80S or 40S ribosomes at initiator AUG codons, and primer extension inhibition (toeprinting) assays were used to delineate the specific positions of those ribosomes on the different mRNAs. Polysome profiling assays, employing sucrose-gradient fractionation of translation reactions and subsequent northern blotting of gradient fractions, were used to assess ribosome–mRNA association. The cap analogue m^7GpppG was used in all experiments as a competitive inhibitor of translation initiation.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 10 February; accepted 9 April 2008.
Published online 21 May 2008.

1. Jacobson, A. in *Translational Control* (eds Hershey, J. W., Mathews, M. B. & Sonenberg, N.) 451–480 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1996).
2. Gallie, D. R. The cap and poly(A) tail function synergistically to regulate mRNA translational efficiency. *Genes Dev.* **5**, 2108–2116 (1991).
3. Wells, S. E., Hillner, P. E., Vale, R. D. & Sachs, A. B. Circularization of mRNA by eukaryotic translation initiation factors. *Mol. Cell* **2**, 135–140 (1998).
4. Dmitriev, S. E., Pisarev, A. V., Rubtsova, M. P., Dunaevsky, Y. E. & Shatsky, I. N. Conversion of 48S translation preinitiation complexes into 80S initiation complexes as revealed by toeprinting. *FEBS Lett.* **533**, 99–104 (2003).
5. Amrani, N. *et al.* A *faux* 3′-UTR promotes aberrant termination and triggers nonsense-mediated mRNA decay. *Nature* **432**, 112–118 (2004).
6. Wu, C., Amrani, N., Jacobson, A. & Sachs, M. S. The use of fungal *in vitro* systems for studying translational regulation. *Methods Enzymol.* **429**, 203–225 (2007).
7. Tarun, S. Z. Jr & Sachs, A. B. A common function for mRNA 5′ and 3′ ends in translation initiation in yeast. *Genes Dev.* **9**, 2997–3007 (1995).
8. De Gregorio, E., Preiss, T. & Hentze, M. W. Translational activation of uncapped mRNAs by the central part of human eIF4G is 5′ end-dependent. *RNA* **4**, 828–836 (1998).
9. Sachs, A. in *Translational Control of Gene Expression* (eds Sonenberg, N., Hershey, J. W. & Mathews, M. B.) 447–465 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2000).
10. Karim, M. M. *et al.* A mechanism of translational repression by competition of Paip2 with eIF4G for poly(A) binding protein (PABP) binding. *Proc. Natl Acad. Sci. USA* **103**, 9494–9499 (2006).
11. Christensen, A. K., Kahn, L. E. & Bourne, C. M. Circular polysomes predominate on the rough endoplasmic reticulum of somatotropes and mammotropes in the rat anterior pituitary. *Am. J. Anat.* **178**, 1–10 (1987).
12. Baer, B. W. & Kornberg, R. D. The protein responsible for the repeating structure of cytoplasmic poly(A)-ribonucleoprotein. *J. Cell Biol.* **96**, 717–721 (1983).
13. Sachs, A. B., Bond, M. W. & Kornberg, R. D. A single gene from yeast for both nuclear and cytoplasmic polyadenylate-binding proteins: domain structure and expression. *Cell* **45**, 827–835 (1986).
14. Sachs, A. B., Davis, R. W. & Kornberg, R. D. A single domain of yeast poly(A)-binding protein is necessary and sufficient for RNA binding and cell viability. *Mol. Cell. Biol.* **7**, 3268–3276 (1987).

15. Gebauer, F., Corona, D. F., Preiss, T., Becker, P. B. & Hentze, M. W. Translational control of dosage compensation in *Drosophila* by Sex-lethal: cooperative silencing via the 5′ and 3′ UTRs of msl-2 mRNA is independent of the poly(A) tail. *EMBO J.* **18,** 6146–6154 (1999).
16. Munroe, D. & Jacobson, A. mRNA poly(A) tail, a 3′ enhancer of translational initiation. *Mol. Cell. Biol.* **10,** 3441–3455 (1990).
17. Otero, L. J., Ashe, M. P. & Sachs, A. B. The yeast poly(A)-binding protein Pab1p stimulates *in vitro* poly(A)-dependent and cap-dependent translation by distinct mechanisms. *EMBO J.* **18,** 3153–3163 (1999).
18. Goyer, C. *et al.* TIF4631 and TIF4632: two yeast genes encoding the high-molecular-weight subunits of the cap-binding protein complex (eukaryotic initiation factor 4F) contain an RNA recognition motif-like sequence and carry out an essential function. *Mol. Cell. Biol.* **13,** 4860–4874 (1993).
19. Kessler, S. H. & Sachs, A. B. RNA recognition motif 2 of yeast Pab1p is required for its functional interaction with eukaryotic translation initiation factor 4G. *Mol. Cell. Biol.* **18,** 51–57 (1998).
20. Cosson, B. *et al.* Poly(A)-binding protein acts in translation termination via eukaryotic release factor 3 interaction and does not influence PSI^+ propagation. *Mol. Cell. Biol.* **22,** 3301–3315 (2002).
21. Mangus, D. A., Evans, M. C. & Jacobson, A. Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression. *Genome Biol.* **4,** 223.1–223.14 (2003).
22. Salas-Marco, J. & Bedwell, D. M. GTP hydrolysis by eRF3 facilitates stop codon decoding during eukaryotic translation termination. *Mol. Cell. Biol.* **24,** 7769–7778 (2004).
23. Stansfield, I., Kushnirov, V. V., Jones, K. M. & Tuite, M. F. A conditional-lethal translation termination defect in a *sup45* mutant of the yeast *Saccharomyces cerevisiae*. *Eur. J. Biochem.* **245,** 557–563 (1997).
24. Uchida, N., Hoshino, S., Imataka, H., Sonenberg, N. & Katada, T. A novel role of the mammalian GSPT/eRF3 associating with poly(A)-binding protein in cap/poly(A)-dependent translation. *J. Biol. Chem.* **277,** 50286–50292 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Kervestin for providing us with recombinant Pab1; J. McCarthy for eIF4E antibodies; A. Hinnebusch for the *cdc33* strain; D. Bedwell for the plasmid-borne *sup35-R419G* allele; and members of the Jacobson laboratory for comments and discussions. This work was supported by a grant to A.J. from the National Institutes of Health.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.J. (allan.jacobson@umassmed.edu).

METHODS

Synthetic mRNAs. Synthetic, capped poly(A)-containing miniUAA1, UAA and AAA mRNAs were transcribed *in vitro* from chimaeric genes cloned in a pSP65A vector that included about 65 dT residues for transcription of a poly(A) tail[5]. mRNAs of intermediate lengths originated from truncated DNA constructs derived from pSP65A-CAN1/LUC[5] by deletion of the *LUC* open reading frame from *Pst*I restriction sites (created by site-directed mutagenesis; Stratagene). mRNAs differing in poly(A) tail length were synthesized from miniUAA1 DNA cloned in pSP65A vectors containing the corresponding lengths of dT residues[16]. The construct used for synthesis of long ADE2 mRNA was generated in a one-step PCR amplification, using *Eco*RI-site-containing oligonucleotide no. 165 (which also contains mutations of two upstream AUGs) (5′-CGGAATTCATTATAGAGCATTTCATATATAAATTGGTGCGTAAAATCGTTGGATCTCTC-3′) and *Bam*HI site-containing oligonucleotide no. 164 (5′-GAGGATCCAAATTCTTAAAAAAGGACACCTGTAAGCGTTG-3′) and cloned into the *Eco*RI and *Bam*HI restriction sites of the pSP65A vector. The miniADE2 DNA construct was generated by deletion of the 1,575-nt fragment from a *Pst*I restriction site 260–265 nt downstream of the start of the construct and a *Pst*I site immediately 5′ to the normal stop codon (both sites were created by site-directed mutagenesis). All plasmid constructions were confirmed by DNA sequencing. Capped and polyadenylated RNAs were synthesized with the SP6 mMessage mMachine kit (Ambion), in accordance with the manufacturer's protocol, from *Hin*dIII-linearized plasmids for poly(A)-containing mRNA or from *Cla*I linearized plasmids for mRNA lacking the poly(A) tail. At least 80% of *in vitro* transcribed mRNAs were capped, as determined by immunoprecipitation with anti-2,2,7-trimethylguanosine agarose conjugate (Calbiochem). Uncapped mRNA was synthesized with the MEGAscript kit (Ambion). RNA yields were quantified by spectrophotometry and their integrities were assessed by agarose-gel electrophoresis.

Extracts and translation reactions. *Saccharomyces cerevisiae* strains[25–29] (Supplementary Table 1) were used to make extracts for *in vitro* translation by techniques described previously[6]. Wild-type strains from different genetic backgrounds gave similar results. Translation and toeprinting assays were essentially the same as those described previously[5]. Unless indicated otherwise, translation reactions were incubated for 4 min with 0.06 pmol of each RNA substrate, and terminated by incubation for 3 min with 2 mM cycloheximide or GMP-PNP. Where indicated, translation initiation was subjected to competitive inhibition by 2.7 mM cap analogue (m^7GpppG), a concentration that readily distinguishes between initiation complexes that are, respectively, sensitive or resistant to this compound. Toeprinting used two different primers that were complementary to their respective targets at comparable distances from the AUG of interest. Primer no. 3029 was used for mRNAs spanning 600 nt (miniUAA2) to 2,135 nt (AAA), and primer no. 55 was used with miniUAA1 mRNA. Samples from translation reactions programmed with ADE2 and miniADE2 mRNAs were toeprinted with primer no. 172, complementary to the sequence 5′-CACTAAAGAATCTTCAAGTAAAACATCCC-3′, and primer no. 3238, complementary to the sequence 5′-GGACTTCATACATAGAAATCAACG-3′, respectively. For mRNA–ribosome association assays, the equivalent of four translation reactions were fractionated in an 11-ml sucrose gradient (7–47%). The gradients were fractionated and scanned at 254 nm, and the resulting absorbance profiles were used to determine the position of the polysomal and non-polysomal fractions[30]. RNA was extracted from each fraction and analysed by northern blotting as described previously[30].

25. Iizuka, N. & Sarnow, P. Translation-competent extracts from *Saccharomyces cerevisiae*: effects of L-A RNA, 5′ cap, and 3′ poly(A) tail on translational efficiency of mRNAs. *Methods* **11**, 353–360 (1997).
26. Mangus, D. A., Amrani, N. & Jacobson, A. Pbp1p, a factor interacting with *Saccharomyces cerevisiae* poly(A)-binding protein, regulates polyadenylation. *Mol. Cell. Biol.* **18**, 7383–7396 (1998).
27. Tarun, S. Z. Jr, Wells, S. E., Deardorff, J. A. & Sachs, A. B. Translation initiation factor eIF4G mediates *in vitro* poly(A) tail-dependent translation. *Proc. Natl Acad. Sci. USA* **94**, 9046–9051 (1997).
28. Barnes, C. A., Singer, R. A. & Johnston, G. C. Yeast *prt1* mutations alter heat-shock gene expression through transcript fragmentation. *EMBO J.* **12**, 3323–3332 (1993).
29. Nielsen, K. H. *et al.* Functions of eIF3 downstream of 48S assembly impact AUG recognition and GCN4 translational control. *EMBO J.* **23**, 1166–1177 (2004).
30. Mangus, D. A. & Jacobson, A. Linking mRNA turnover and translation: assessing the polyribosomal association of mRNA decay factors and degradative intermediates. *Methods* **17**, 28–37 (1999).

naturejobs

JOBS OF THE WEEK

College and graduate-level students need better preparation for jobs in industry — that was a major theme at sessions on the biotech workforce at the Biotechnology Industry Organization's convention in San Diego, California, last week.

Speakers touted various remedies. Sheldon Schuster, president of the Keck Graduate Institute in Claremont, California, pointed to the success of the institute's professional master's programme, which combines science coursework with MBA-level classes in business management and finance. Schuster claimed that 97% of Keck graduates found industry jobs within six months of getting their degree. David Cheresh of the Moores Cancer Center at the University of California, San Diego described the UC Discovery Fellowship, which funds positions that facilitate relationships between industry and academia. One recent fellow set up an initiative on his campus to support entrepreneurs with help on consulting, networking and finding seed funding.

Some emphasized the training role that community colleges could play. In the United States, four-year degrees — even from prestigious schools — often do not provide the practical lab skills needed for entry-level industry jobs, according to Elaine Johnson, director of the Bio-Link National Center in San Francisco, which fosters curriculum improvement and professional development at community colleges and elsewhere. Community colleges could fill the breach.

Industry needs to help with on-site initiatives as well. South San Francisco-based Genentech, for example, has 100 postdocs on short-term contracts in-house. There is no promise of a permanent post at the end of their contract, says Genentech's David Chang. The rationale, he explains, is that postdocs who think a job is on the line are more likely to take a conservative approach to their research. Genentech's programme is not entirely altruistic — the company gets an occasional hiring and constructs a network of postdoc alumni. But it, like the above initiatives, recognizes that the success of biotech as a whole has more to do with people and talent than with equipment or buildings. ■

Gene Russo is editor of Naturejobs

CONTACTS
Editor: Gene Russo

European Head Office, London
The Macmillan Building,
4 Crinan Street, London N1 9XW, UK
Tel: +44 (0) 20 7843 4961
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

European Sales Manager:
Andy Douglas (4975)
e-mail: a.douglas@nature.com
Business Development Manager:
Amelie Pequignot (4974)
e-mail: a.pequignot@nature.com
***Nature*events:**
Claudia Paulsen Young (+44 (0) 20 7014 4015)
e-mail: c.paulsenyoung@nature.com
France/Switzerland/Belgium:
Muriel Lestringuez (4994)
Southwest UK/RoW: Nils Moeller (4953)

Scandinavia/Spain/Portugal/Italy:
Evelina Rubio-Hakansson (4973)
Northeast UK/Ireland:
Matthew Ward (+44 (0) 20 7014 4059)
North Germany/The Netherlands:
Reya Silao (4970)
South Germany/Austria:
Hildi Rowland (+44 (0) 20 7014 4084)

Advertising Production Manager:
Stephen Russell
To send materials use London address above.
Tel: +44 (0) 20 7843 4816
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com
***Naturejobs* web development:** Tom Hancock
***Naturejobs* online production:** Dennis Chu

US Head Office, New York
75 Varick Street, 9th Floor,
New York, NY 10013-1917
Tel: +1 800 989 7718
Fax: +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

India
Vikas Chawla (+91 1242881057)
e-mail: v.chawla@nature.com

Japan Head Office, Tokyo
Chiyoda Building, 2-37 Ichigayatamachi,
Shinjuku-ku, Tokyo 162-0843
Tel: +81 3 3267 8751
Fax: +81 3 3267 8746

Asia-Pacific Sales Manager:
Ayako Watanabe (+81 3 3267 8765)
e-mail: a.watanabe@natureasia.com
Business Development Manager, Greater China/Singapore:
Gloria To (+852 2811 7191)
e-mail: g.to@natureasia.com

NATURE|Vol 453|26 June 2008

MOVERS

James Halpert, Associate Dean for Scientific Affairs, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego

2003–2008 Director, Environmental Health Sciences Center, University of Texas Medical Branch, Galveston, Texas
1998–2008 Professor and Chairman, Department of Pharmacology and Toxicology, University of Texas Medical Branch, Galveston, Texas

James Halpert is a self-described cautious, methodical guy. But his career path suggests a bit of daring. He left promising undergraduate work in chemistry to travel to Europe and learn a new language. And he has uprooted himself repeatedly when eager for new challenges, a tendency that he says has benefited his scientific career.

After receiving a bachelor's degree in Scandinavian languages from the University of California, Los Angeles, he took up laboratory work at a Swedish hospital, ultimately deciding to pursue biochemistry. He deciphered the amino acid sequence of a deadly snake venom neurotoxin while earning his PhD at Uppsala University. Having published several papers on natural toxins, Halpert's interest shifted to manmade toxins; he went on earn an MSc in toxicology at the Karolinska Institute.

After seven years in Sweden, Halpert returned to the United States and a postdoc at Vanderbilt University in Nashville, Tennessee. It was there that he began work on the cytochrome P450 superfamily, the most important element of drug metabolism and a focus that would tie together his interests in biochemistry, the environment and human health. Halpert was the first to determine how various P450s are inhibited while carrying out reactions to metabolize drugs.

"Jim's work is always at the cutting edge," says Paul Hollenberg, a pharmacologist at the University of Michigan.

Shortly after joining the University of Arizona — where he became a professor of pharmacology and deputy director of the Southwest Environmental Health Sciences Center— the junior faculty member made an important decision. "At 37, I was worried I might become a dinosaur if I didn't learn molecular biology techniques," he says.

A move to the University of Texas Medical Branch at Galveston allowed Halpert to focus on structural biology and do some of the best work of his career — solving several P450 structures. Eventually, he became the director of the university's National Institute of Environmental Health Sciences Center.

As the new associate dean for pharmaceutical sciences at the University of California at San Diego, Halpert plans to recruit a cadre of researchers and move beyond simply training pharmacists, in part by a nascent joint PharmD/PhD programme. "We want to create researchers able to develop the next generation of drugs," he says. ■
Virginia Gewin

NETWORKS & SUPPORT

Postdoc competencies

In 2004, the US National Postdoctoral Association (NPA) set out to identify and recommend postdoctoral 'best practices', a document the NPA issued in 2005. We suggested things such as standardized classification, establishing postdoctoral offices and implementing specific policies regulating appointments.

In releasing these best practices, the NPA asked research institutions to develop specific competencies crucial to a successful training experience. In response, institutions challenged us to develop guidelines, a draft of which was presented at the NPA annual meeting held in April. We believed that the dramatic change in the number and demographics of US postdocs made targeted training vital. And we expect these competencies will help postdocs and their mentors in developing rational career plans.

With input from postdocs, faculty and administrators, the NPA policy committee created a set of core principles that was recently released. We did not try to develop a one-size-fits-all solution. Rather, we sought to identify a set of skills flexible enough to let institutions with different populations and resources develop their own programmes. The final competencies (scientific knowledge, research skills, communication skills, professionalism, leadership and management skills, and responsible conduct of research) cover all career outcomes and research fields.

This list raised the question: how should the competencies be evaluated and regulated, if at all? Mandatory evaluation would be misplaced, as postdoctoral positions, programmes and their objectives vary tremendously among institutions.

They are simply a tool to help improve and customize training. Institutions with a uniform method for classifying postdoctoral positions may see a valuable role for regulation. Others may develop resources such as short courses, seminars or dedicated professional development assistance. And postdocs can incorporate the principles into individual development plans, a 'mapping out' of expectations done in conjunction with one's adviser.

We are planning to release the core competencies as a Web-based resource in autumn 2008. To finalize the document, we would appreciate feedback, including resources that could inform the competencies.
Lisa Curtis, Keith Micoli and Jennifer Reineke Pohlhaus are current and former members of the board of directors of the National Postdoctoral Association.

POSTDOC JOURNAL

Cool and collected

"I thought of ice-cream today," my fellow researcher announced last night. We were sitting outside, shivering in three layers of clothing, clutching beers to toast the setting Ethiopian sun. I nodded as he added that the fat little birds called francolins that run in front of the truck every day make him salivate. They look so plump and delicious. I can relate. We are far from starving, but food has become the topic of every second conversation. Even the grass looks tasty.

Being out in the middle of nowhere with minimal comforts can transform trained intellectuals into beings whose mood depends on the amount of rain that fell the night before. It becomes hard to look past the here and now; cravings for our favourite foods tie us to the world beyond the Simien Mountains. Little victories help us to adjust, from a successful two-minute conversation in Amharic (the local language) to making pancakes that taste like the real thing.

For a couple of weeks I've had the company of two fellow researchers: a graduate student and a professor out here to investigate gelada behaviour. After three solitary months it was wonderful to engage in intellectual conversation, babble about ideas and experiments and politics. But after a spell, the focus changed to basic cravings. I really just want to know how to get my hands on a T-bone steak, right now. I'll trade my field hat for it.
Aliza le Roux is a postdoctoral fellow in animal behaviour at the University of Michigan.

Dead yellow

Hue and cry.

Tanith Lee

This was my wedding dress. At the time people remarked on my choice of colour, but with my hair the way I had it then, it worked. I remember there were daffodils blooming. But I won't show you the photographs. No point now, is there?

When did it start? Officially in 2036. But the papers had been reporting curious anomalies for years before that. And people spotting things. Thinking at first the fault was in them and getting frightened — so many medical case-notes.

And I? Oh, I think I first properly *noticed* that day when we walked in the park. We often did that, then. It was a nice park, lots of trees, wild areas. But I heard a child — it's funny, isn't it, the way children always ask the truly awful question? — this child said to some adult, "Why are all the trees going brown?" And it was late May, you understand, early summer, and the leaves flooding out and the grass high and everything lush. What did the adult reply? I can't recall. But as we walked on, the scales, as they say, dropped from my eyes. I wish they hadn't. I began to see it too.

It wasn't like it is nowadays. Then it was only just establishing itself, the — what did they call it? — The *Phenomenon*.

It was almost like looking through a photographic lens. Except, obviously, this lens didn't completely change everything, as normally it would.

Neither of us said anything to the other. But I realized he, my husband, had also in those moments begun to see. We kept talking and joking, we even stopped for coffee and a doughnut at the park café. But an uneasy shadow was settling on us, and a silence.

We didn't actually discuss anything for several weeks. One evening we were making dinner, and — I remember so vividly — he was suddenly staring at the counter and he said, "What colour is that pepper, would you say?"

"Sort of orange, I suppose," I said, "an orange pepper."

"No," he said, "it's a brown pepper. And the lettuce, that's a pale brown lettuce, only its edges are … pale blue."

And we had become two statues, while the cooker bubbled carelessly, and then he said, "Someone at work went for his eye-test today. He'd told me he was afraid he was going blind. But his problem isn't caused by any defect in his vision. The optician said, apparently, the problem is becoming universal."

JACEY

And then, as if we must, we looked around us, at all and everything: the brown curtains that had been a deep green, and the green trees beyond the windows that were the colour of sludge, yes, even in the evening light — where the blue sky was somehow wrong and the west such a dark and sullen red. In the clear glass bottle the white wine gleamed colourless as water, but the mustard in the jar was mud. And on my hand my gold wedding ring had altered to the dull metal of a tarnished, ancient penny.

"What is it?" I said.

"God knows," he said. But I don't think God, if there is God, does know either, any more than the rest of us.

We all comprehend by now, or I assume most of us must do. It's worldwide after all. Hardly anyone talks about it. Aside from very young people like yourself, who never watched it happen. It's meant a lot of make-overs, home décor, clothing — good for commerce then. Even I had my corn-blonde hair bleached dead white. Better than the stagnant-pond shade it had become. (Like my wedding dress, as you see.) And if no one wants black-brown-blue cabbages and lettuces, or eggs with blind-brown centres, or the quite fresh yet decayed-looking brownish peaches and apricots, there are still things to eat. Apples and tomatoes like an old wound, doughnuts like excrement. The jewel trade suffered. Who buys a topaz? A cut emerald the size of a cat's (brown/grey) eye, is worth less than nine eurodollars — less than the price of a bottle of good (stale-tea colour) Pinot Grigio. Or black Merlot.

It's worse for animals. Those white leopards that lost their camouflage, the brown canaries that stopped breeding and died out — as the leopards and the tigers did. And overhead the Sun is molten white or murky crimson, and the Moon ashes, that sometimes curdle into blood.

Because yellow was a primary colour it didn't die alone. It took green and orange with it, and virtually every other shade lost some nuance or definition. How strange. Who could ever have guessed? They said that some kind of spectrum-microbe caused this. It attacked only that one element, the colour yellow. Nothing dangerous, no need for alarm, can't harm us. Just … hurts. No, I won't show you the photos. It affects photographs too, of course. That girl in a *brown* dress, the *brown* and *bone-white* daffodils …

My husband? I'm afraid he died young.

Thank you for your visit. Yes, isn't it a dramatic sunset?

Apocalyptic, you could say. ■

Tanith Lee: written 97 books, 261 short stories, four radio plays and two episodes of science-fiction series *Blake's 7*. Is married. Has cats.